GeneSpring
User Manual

version 4.1

Release date, 27 September 2001

Copyright 1998-2001 Silicon Genetics. All rights reserved. GeneSpring, GeneSpider, GenEx, GeNet, and MicroSift are trademarks of Silicon Genetics. All other products, including but not limited to Affymetrix GeneChip®, Affymetrix Global Scaling™, GenBank, Microsoft Excel®, Microsoft Notepad® and Adobe FrameMaker®, are the trademarks of their respective holders.

Related Documents

GeneSpring Basics Instructional Manual, version 4.0.2. Release date, 31 May 2001

GeNet User Manual, version 2.3. Release date, 12 June 2001

Table of Contents

Chapter 1 Introduction 1-1
Getting Started 1-1
Learning to Use GeneSpring 1-3
New in Version 4.0 1-4
GeneSpring Basics 1-7
The GeneSpring Hierarchy of Objects or, Where Is My Data Stored? 1-15
Commonly Used GeneSpring Functions 1-17
The Gene Inspector window 1-17
Making Lists 1-17

Chapter 2 Creating DataObjects in GeneSpring 2-1
The Experiment Autoloader 2-1
Autoloader Normalizations 2-3
Default Normalizations of Commercially Available Products 2-4
Merging, Splitting and Duplicating Experiments 2-6
Loading from Subchips 2-7
Creating a Genome through the Autoloader 2-7
Change Experiment Parameters 2-8
The Experiment Parameters Window 2-9
Add a Parameter 2-10
Re-order the Parameters 2-10
Definitions of Parameters 2-11
Parameter Vocabulary 2-11
Parameters Displayed in the Navigator 2-11
A Note on Multiple Parameters 2-12
Parameter Display Options 2-12
Continuous Element 2-13
Non-Continuous Element (Set) 2-13
Color Code 2-13
Annotation Tools 2-15
Updating your Master Gene Table with GeneSpider 2-15
Building a Simplified Ontology 2-16
Changing the Experiment Interpretation 2-17
Vertical Axis Modes 2-18
Parameter Display Modes 2-20
Experiment Normalizations 2-21
Background Subtraction 2-21
Per-spot Normalization 2-22

Copyright 2000-2001 Silicon Genetics

Per-chip Normalizations 2-22
Use Positive Control Genes 2-22
Normalizing to the Distribution of All Genes 2-23
Region Normalization 2-23
The Affine Background Correction 2-23
Use Constant Values 2-24
Per-gene Normalizations 2-25
Normalize to Median For Each Gene 2-25
Normalizing to Sample(s) 2-25
Miscellaneous 2-26
Global Error Models 2-26
Using the Global Error Model 2-26
Technical Details 2-28

Chapter 3 Viewing Data in GeneSpring 3-1
Using Genome Browser 3-1
Changing Genome Browser Elements 3-2
Splitting Windows 3-3
Displaying a Gene List 3-4
Finding and Selecting Genes 3-4
Finding Genes 3-4
Selecting Genes 3-5
Showing/Hiding Window Display Elements 3-6
Graph View 3-7
Bar Graph View 3-8
Classifications View 3-9
Physical Position View 3-10
Scatter Plot View 3-15
Tree View 3-17
Magnifying Trees 3-18
Selecting and Viewing Subtrees 3-18
Viewing Nodes 3-18
Viewing Gene Names in Trees 3-19
Viewing Colors in Trees 3-19
Viewing Parameters in Trees 3-19
Horizontal Genes/Vertical Genes 3-20
Ordered List View 3-21
Array Layout View 3-22
Pathway View 3-23
Compare Genes to Genes 3-24
Graph by Genes View 3-26
Functional Classification 3-27
View as Spreadsheet 3-29
Linked Windows 3-30
Split Windows 3-30
Bookmarks 3-31
Changing the Coloring Scheme 3-31

 Copyright 2000-2001 Silicon Genetics

Color by Expression 3-31
Color by Significance 3-33
Color by Static Experiment 3-33
Color by Venn Diagram 3-33
Color by Parameter 3-33
No Color 3-34
Color by Classification 3-34
Color by Secondary Experiment 3-35
Changing the Experimental Data Range 3-36
Changing the Default Colors 3-37
The Inspectors 3-37
Gene Inspector 3-37
Experiment and Condition Inspectors 3-41
Condition Inspector 3-43
List Inspector 3-44
Classification Inspector 3-46

Chapter 4 Analyzing Data in GeneSpring 4-1
Filter Genes Analysis Tools 4-1
Restrictions Over an Entire Experiment or Interpretation 4-3
Restrictions over a Single Condition or Sample 4-7
Restricting by Associated Numbers 4-9
New Gene List window 4-11
Making Lists with the Find Similar Command 4-13
Making Lists with the Complex Correlation Command 4-14
The Multi-Experiment Correlation Window 4-15
Finding Offset Genes 4-18
Making Lists from Properties 4-19
Making Lists with the Venn Diagram 4-19
Making Lists from Classifications 4-21
Find Interesting Genes 4-21
Making Lists from Selected Genes 4-22
Creating Drawn Genes 4-22
Pathways 4-23
Importing a Pathway 4-24
Adding a Gene to a Pathway 4-24
Adding KEGG Pathways 4-25
Finding New Genes on a Pathway 4-25
Regulatory Sequences 4-26
Making Lists of Homologs and Orthologs 4-31
Scripts 4-32
Using Scripts 4-32
What is a Script? 4-32
Creating Your own Scripts 4-34
Auto-Publish to GeNet 4-40

 Copyright 2000-2001 Silicon Genetics

External Programs 4-40
GeneSpring External Program Interface 4-40
Examples 4-42

Chapter 5 Clustering and Characterizing Data in GeneSpring 5-1
Trees 5-1
Creating a New Gene Tree 5-1
Creating Complex Experiment Trees 5-2
References for Hierarchical Clustering 5-4
Principal Components Analysis 5-5
References for Principal Components Analysis 5-8
k-Means Clustering 5-9
Viewing k-means clusters 5-11
Self-Organizing Maps 5-12
Viewing SOMs 5-13
The Class Predictor 5-15
Interpreting the Results of a Prediction 5-16

Chapter 6 Exporting GeneSpring Data 6-1
Saving Pictures and Printing 6-2
Exporting Gene Lists out of GeneSpring 6-3
Publish to GeNet 6-6
Upload to GeNet 6-6
Using GeNet 6-8
Loading Data from GeNet 6-8

Appendix A Help A-1
Contacting Silicon Genetics' Technical Support A-1
The Help Menu A-1
GeneSpring Basics Instructional Manual A-1
Manual A-1
FAQ A-1
Version Notes A-1
Update GeneSpring A-2
Silicon Genetics on the Web A-2
GeNet Database A-2
Register for a Workshop A-2
System Monitor A-2
About A-2

 Copyright 2000-2001 Silicon Genetics

Appendix B Preferences Window B-1
Data Files B-1
Database B-1
Color B-2
Specific Color Definition B-3
Gene Labels B-4
Browser Details B-4
The Firewall Details box B-4
The System Preferences B-5
The Miscellaneous B-5

Appendix C Genome Wizard C-1

Appendix D The Experiment Wizard D-1
Files You will Need to Use the Experiment Wizard D-1
The Experiment Import Wizard D-3

Appendix E Installing from a Database E-1
Custom Databases and GeneSpring E-1
Databases E-1
Open Database Connectivity E-1
Structured Query Language E-2
SQL Call Level Interfaces E-2
The Genetic Analysis Technology Consortium E-2
Databases and GeneSpring E-3
Adding an Experiment from a Database E-3
Test to Make Sure Your ODBC Connection is Working E-4
Connect your Database to GeneSpring E-4
Entering your Prepared Database into GeneSpring E-5
Entering more Complicated Data from a Database E-6

Appendix F Copying and Pasting Experiments F-1
Preparation for Pasting F-1
Most Common Mistakes in Pasting F-3
Pasting your Experiment into GeneSpring F-4
Copying an Experiment or a List Out of GeneSpring F-4

Appendix G Normalizing Options G-1
Background Subtractions G-2
Normalize to Negative Controls G-2
Mathematical Illustration of the Normalize to Negative Controls Method G-2
Normalize to Control Channel Values for Each Gene G-3
Mathematical Illustration of the Normalize to a Control Channel Value for Each Gene Method G-4

 Copyright 2000-2001 Silicon Genetics

Normalize to Positive Controls G-5
Mathematical Illustration the Normalize to Positive Controls Method G-5
Normalize Each Sample to Itself G-6
Mathematical Illustration of the Normalize Each Sample to Itself Method G-6
Normalizing Each Sample to a Hard Number G-7
Normalizing Each Gene to Itself G-8
Mathematical Illustration of the Normalizing Each Gene to Itself Method G-8
Normalizing All Samples to Specific Samples G-10
Required Syntax for Normalization to Specific Samples G-10
Mathematical Illustration of the Normalizing Samples to a Specific Sample Method G-12
Region Normalization G-15
Dealing with Repeated Measurements G-16
Single Data File G-16
Mathematical Illustration of the Dealing with Repeated Measurements in a Single Data File Method G-16
Measurement Flags G-17
Negative Control Strengths G-18
Normalization for Particular Array Types G-18

Appendix H Creating Folders for New Genomes H-1
Raw Data H-1
What Data Are Necessary? H-1
What Format do these Data Need to be in? H-1

Appendix I Installing a Genome from a Text File I-1
Creating Folders for New Genomes I-1
The *.genomedef* File I-1
Define Your Genome I-2

Appendix J Installing from a Text File J-1
Define Your Experiment J-1
Define Your Parameters J-2
Describe your Data Files J-6
Data File Header Lines J-7
Gene Names J-8
Explain to GeneSpring how to locate only the Gene Name J-8
Explain to GeneSpring How to Read the Region Specifications J-9
The required *.layout* file for Region Specifications J-9
Locate the Data Column J-9
The Control Channel Value J-11
Measurement Flags J-12
Associating a Picture with a Sample J-13
Normalizations: Negative Controls J-14
The required layout file for negative controls J-15
Normalizations: Control Channel Values J-15

 Copyright 2000-2001 Silicon Genetics

Normalizations: Positive Controls J-16
The required layout file for positive controls J-16
Normalizations: Each Sample to Itself J-17
Normalizations: Each Gene to Itself J-18
Normalizations: Each Sample to a Specific Sample J-18
Colorbar Specifications J-19
Graph Specifications J-19

Appendix K Experiment File Formats K-1
Raw Data K-1
What format does this data need to be in? K-2
Experimental Data K-2
Pictures of the conditions during the experiment K-2
Pictures of the Microarray plates K-2
The Layout file K-2
The Region Designation File(s) K-4
Entering region specifications when they are not specified in their own column or as suffixes within another column K-5
How to describe a map K-7
The Positive and Negative Control Files K-7
Where do I put my data? K-8

Appendix L Equations for Correlations and other Similarity Measures L-1
Common Correlations L-2
Standard Correlation L-2
Pearson Correlation L-2
Spearman Correlation L-3
Spearman Confidence L-3
Two-sided Spearman Confidence L-3
Distance L-4
Special Case Correlations L-4
Smooth Correlation L-4
Change Correlation L-5
Upregulated Correlation L-5

Appendix M Creating an Array in GeneSpring M-1
Examples of *.layout* files for Arrays M-2

Appendix N Technical Details on the Statistical Group Comparison N-1
For Each Gene N-1
References N-4

 Copyright 2000-2001 Silicon Genetics

Appendix O Technical Details for the Predictor O-1
Gene Selection O-1
Classifying the Test Samples O-1
Decision Threshold O-1
References for the Predictor O-2

Appendix P Common Commands P-1
Commands Accessible by Cursor or Keyboard P-1
Common Commands in the Drop-Down menus P-2
The File Menu P-2
The Edit Menu P-2
The View Menu P-3
The Experiments Menu P-3
The Colorbar Menu P-3
The Tools Menu P-4
Common Commands in the Genome Browser P-5
The Options Submenu P-5
The Error Bars Submenu P-7
Common Commands in the Navigator P-7
The Main Folder Pop-up Menus P-8
The Gene Lists Folders Pop-up Menus P-8
Common Commands in the Experiment Specification area P-10

Appendix Q Glossary Q-1

Index 1-1

 Copyright 2000-2001 Silicon Genetics

Chapter 1 Introduction

Welcome to GeneSpring. Congratulations on selecting the most advanced, flexible tool available for gene expression data analysis.

This manual is a guide to GeneSpring features. To see the many features new to version 4.1, see “New in Version 4.0” on page 1-4. Chapter 1 will cover installing GeneSpring, loading and setting up your data, and GeneSpring basics. The remaining chapters will discuss loading, set-up and the various data analysis and visualization tools in detail.

Getting Started

Requirements

- A computer with 128 MB RAM (256 MB strongly recommended) with a Pentium II, Celeron, PowerPC, or faster processor.
- Approximately 130 MB including documentation.
- The recommended screen resolution is 1024x768 with a minimum of 16 bit color.

Installing from a CD

If you are installing GeneSpring from a CD, you will see several options after you place your CD in the drive:

1. Select **`Install GeneSpring Demo`**. A splash screen and an Install Anywhere© screen will appear with a progress bar.
2. Follow the on-screen instructions. For more information see the ReadMe file included with the CD.

In Windows, you can also install the software by using the **`Start > Run`** command in the Start menu.

Installing from the Web

If you are reading this manual and do not have a copy of GeneSpring, you can download a copy by going to the following url:
http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/download.smf

Follow the on-screen directions and Silicon Genetics will send you a username, password and download link.

Starting GeneSpring

Once you have installed GeneSpring, you will find two new items on your desktop—the GeneSpring Data folder and the GeneSpring icon.

Copyright 1998-2001 Silicon Genetics

Figure 1-1 The GeneSpring Data and Start icons

To start GeneSpring, double-click the GeneSpring icon. Alternatively, Windows users can reach the GeneSpring icon by selecting Start/Programs/GeneSpring or Program files/Silicon Genetics/GeneSpring. Mac users can also start GeneSpring from the Applications folder/Silicon Genetics/GeneSpring.

A splash screen will appear containing your GeneSpring version number, the expected expiration date and the JVM you are using. You will then see the GeneSpring main window. For further details, see "GeneSpring Basics" on page 1-7.

Obtaining a License Key

If you have already installed a demo copy of GeneSpring, your license key will expire within two months. Once you have purchased a full GeneSpring license, Silicon Genetics will send you a license key. Save this license key file in the Silicon Genetics/GeneSpring/Data folder. (See "The GeneSpring Hierarchy of Objects or, Where Is My Data Stored?" on page 1-15 for details.) On a Windows machine this will be found in C:// Program Files, on a Mac in the Applications folder. When the key is about to expire, you will get a warning message 30 days in advance. If your license has expired or is about to, please contact Silicon Genetics at 866 SIG SOFT (744-7638).

Setting Memory Usage Options

Once GeneSpring is installed, you will need to make sure the default memory setting in GeneSpring preferences is half of your computer's available memory (or more if you have lots of RAM). To do this, select **`Edit > Preferences`**, choose **`System`** from the pull-down menu and enter the amount of memory in the **`Desired Memory Use`** field.

Configuring Virtual Memory (on your hard drive)

Generally, the minimum recommended amount to have available as virtual memory is 150MB RAM. Check to make sure large files are not restricting programs from running as quickly as they might. You may be able to move some large files to another drive.

If you are using the IBM JVM, make sure you specify in the path the appropriate amount of memory to use. You can reach the path by right-clicking the GeneSpring icon on your desktop and choosing **`Properties`** from the pop-up menu. The MS JVM (and the Macintosh JRE) is set to use more of the available memory, but the IBM JVM will as a default use 64MB RAM. For instance, the path specified for the (...java.exe -classpath...) should be changed to include a memory amount equal to about half the RAM on your computer:

```
C:\WINNT\java.exe /cp "D:\Program Files\SiliconGenetics\Gene-
Spring\bin\GeneSpring.jar" GeneSpringMain
```

to

 Copyright 1998-2001 Silicon Genetics

```
C:\WINNT\java.exe -mx164m /cp "D:\Program Files\SiliconGenet-
ics\GeneSpring\bin\GeneSpring.jar" GeneSpringMain
```

If you are still experiencing slowdowns, check the memory usage by selecting **`Help > System Monitor`** before invoking any functions. Make a record of the Total Memory and Free Memory listed in the System Monitor window and contact Silicon Genetics' Technical Services Department at 650-SIG-SOFT or support@sigenetics.com.

Updating GeneSpring

If you already have GeneSpring and just need to obtain the latest update, select **`Help > Update`** and follow the on-screen instructions to obtain the current *GeneSpring.jar*.

Learning to Use GeneSpring

Silicon Genetics provides a variety of ways to improve your knowledge of GeneSpring. In addition to this manual, there is online help, Flash tutorials, a PDF tutorial, and face-to-face workshops that cater to beginning, intermediate or advanced users.

Where to find help

Workshops	http://www.sigenetics.com/cgi/SiG.cgi/Support/workshops.smf
Flash tutorials	http://www.sigenetics.com/cgi/SiG.cgi/Demos/tut_welcome.smf
Tech notes	http://www.sigenetics.com/cgi/SiG.cgi/Documentation/GSTN.smf
FAQs	http://www.sigenetics.com/cgi/SiG.cgi/Documentation/GSFAQ.smf
GeneSpring Tutorial	Go to **`Help > Tutorial`**.
Help buttons on GeneSpring windows	Clicking a **`Help`** button in a given window in GeneSpring opens a page explaining the features of that window.
Technical support	Call Silicon Genetics toll-free at 1 866 SIG SOFT (7638)

Copyright 1998-2001 Silicon Genetics

New in Version 4.0

Scripting

GeneSpring 4.1 can execute scripts to automate data analysis. Users connected to GeNet have the option of running scripts on a remote server.

Easier Data Loading

With just a few clicks of the mouse Gene Spring's new Autoloader makes every attempt to recognize the format of your file and the genome to which it corresponds. If the Autoloader is unfamiliar with your file format, you can use the *Column Editor* to specify the type of data in each column. Once the *Column Editor* learns the location and identity of the relevant columns of data, it adds these specifications to its list of known file types so that you can load subsequent experiments in batch.

The Autoloader now automatically recognizes the following formats:

- Clontech one-color
- Clontech two-color
- Quantarray
- Scanarray4000
- Affymetrix Metrixs
- Affymetrix Pivot
- Axon GenePix 4000
- BioDiscovery Imagene 4
- Incyte Internet
- Incyte GEM Tools 2.4
- Generic one-color
- Generic two-color

Simplified Gene Ontology Construction

The *Build Simplified Ontology* option constructs a simple gene ontology based on keywords from annotations in public databases. The classification scheme is derived from Gene Ontology consortium gene lists. Additional functional classifications were constructed by Silicon Genetics.

Global Error Models

Using the Global Error Model allows you to produce a better estimate of precision. You can use these estimates in a number of analyses in GeneSpring, including filtering and clustering.

Copyright 1998-2001 Silicon Genetics

Statistical Group Comparison

You have three options when choosing Statistical Group Comparison.

- Parametric test, assume variances equal (Student's t-test/ANOVA)
- Parametric test, don't assume variances equal (Welch t-test/Welch ANOVA)
- Non-parametric test (Wilcoxon-Mann-Whitney test/Kruskal-Wallis test)

Class Predictor

The Class Predictor feature allows you to predict the value, or "class", of an individual parameter in an uncharacterized set of samples using a training set where the parameter values are known.

New Inspectors

You can now view at a glance all the data for a particular experiment, condition, interpretation, and classification.

Include Attachments

You can now attach any sort of file to a gene list, experiment, or classification.

Merge/Split Experiments

You can now merge experiments or individual conditions and split experiments.

Customized Clustering Annotations

GeneSpring 4.1 allows the user to define a "standard" group of gene lists to label the branches of a gene expression tree.

Improved Normalization

New on-the-fly normalizations include more robust handling of per-spot normalization, normalization of a region of a chip, and normalization of SAGE data. Also, improved text descriptions of normalization procedures are included in the Interpretation Inspector available for every interpretation.

More Advanced Regulatory Sequence Searching

The *Find Potential Regulatory Sequences* algorithm is now speedier, more flexible, and allows for gaps in the putative consensus sequence.

Copyright 1998-2001 Silicon Genetics

Spreadsheet Display

The Spreadsheet view allows for easy tabular display of expression data for an entire gene list, including:

- normalized signal
- control signal
- raw signal
- t-test p-value
- associated flags

Enhanced Color Options

Expanded color scheme makes visualization of up- and down-regulated genes easier.

Helpful Hints

Helpful hints pop-up dialog boxes will guide you through the data loading process. Also new-and-improved Help buttons appear on many screens throughout GeneSpring.

Copyright 1998-2001 Silicon Genetics

GeneSpring Basics

GeneSpring is a remarkably powerful analysis tool and like any professional level program, it can be intimidating to new users. The following section is a brief introduction to using GeneSpring and loading data, designed to get you up and running in the shortest possible time. Figure 1-2 depicts the steps in a typical analysis session using GeneSpring. Note that this diagram represents what might occur in a typical data analysis session and does not include all of the types of analyses found in GeneSpring.

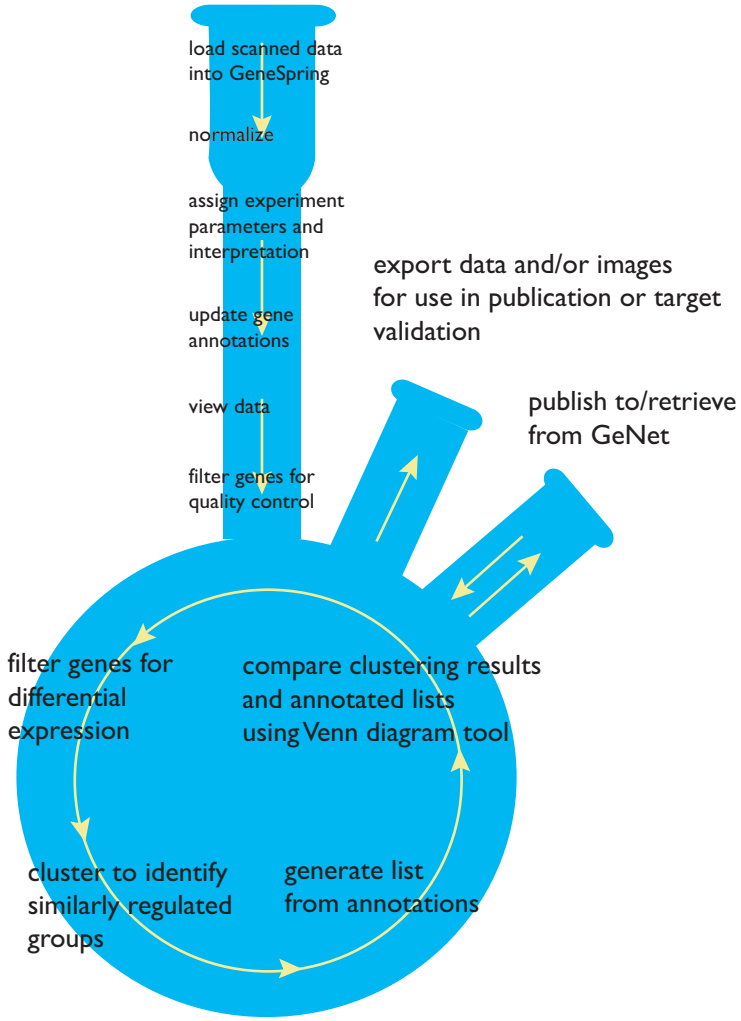

load scanned data into GeneSpring
normalize
assign experiment parameters and interpretation
update gene annotations
view data
filter genes for quality control
export data and/or images for use in publication or target validation
publish to/retrieve from GeNet
filter genes for differential expression
compare clustering results and annotated lists using Venn diagram tool
cluster to identify similarly regulated groups
generate list from annotations

Figure 1-2 Typical GeneSpring workflow

In loading your data, you will come across terms and concepts such as genome, parameter, parameter values, replicate, interpreted data, etc. Below are explanations of how these terms are used in GeneSpring.

Copyright 1998-2001 Silicon Genetics

What is meant by a Genome?

A genome contains information about all the genes in your chip or microarray setup. Note that a GeneSpring genome does not correspond exactly to the biological definition of a genome. A genome in GeneSpring is composed of discrete genes as opposed to the full nucleotide sequence. This means that a GeneSpring genome can contain two genes representing alternatively spliced variants of a single gene, whereas a true genome would only include the DNA sequences for one.

What is meant by a Parameter?

Parameters are experiment variables, such as stage, time, concentration, etc.

Parameter values are values assigned to experiment parameters. For example Embryonic, Postnatal or Adult could be parameter values of the experiment parameter *stage*, while .01 ppm could be a parameter value of the experiment parameter *concentration*.

What is meant by Replicates?

Replicates can be multiple spots on the same array representing the same gene (also referred to as a copy), the same sample on more than one array or a biological replicate—that is equivalent samples taken from more than one organism. Graphically, a parameter defined as a replicate is a hidden variable; no visual distinction is made based upon this parameter or its parameter values.

What is meant by Raw Data?

The analysis process begins by obtaining data in the form of flat files that were generated by your scanning software or other expression analysis technology. GeneSpring is capable of recognizing most commercially available formats and can learn to recognize initially unfamiliar formats as they arise. Typically, the gene/spot/probe-set intensity values in these files are referred to as raw data.

What is meant by Normalized Data?

If GeneSpring recognizes your file format, it will apply a set of default normalizations appropriate for your expression analysis technology. The denominator used to normalize each measurement is referred to as the control strength.

What is meant by Interpreted Data?

GeneSpring is able to interpret normalized data in many different ways. You can elect to have multiple samples treated as replicates and averaged and indicate what type of assumptions you would like GeneSpring to make about the precision of these averaged values. You can display and perform analyses on the normalized data using three modes: ratio (raw versus control strength), logarithm of ratio, or in terms of fold change (versus the control strength). *It is important to note that the graphical display of normalized values and the numbers used for all analyses (such as clustering) reflect the mode you have chosen. However, the numbers displayed as text (as in the Gene Inspector window) and entered by the user as parameters for analyses (as in the Filter Genes tools) are always in ratio mode.*

Copyright 1998-2001 Silicon Genetics

Loading Your Data

The demonstration version of GeneSpring comes pre-loaded with sample yeast, rat and human data. Many users benefit from performing trial analyses on these sample data sets. When you are ready to analyze your own data, you will need to load and set up your data for analysis. There are four main steps to preparing data:

1. Loading gene information (optional).
2. Loading experiment information.
3. Telling GeneSpring how to interpret the information by assigning normalizations, parameter values, and modes of display.
4. Annotating/updating your genome.

To Load Your Data

- **Step 1:** Load gene information from your arrays (optional)

  a. Start GeneSpring and select **`File > New Genome Installation Wizard`**.

  b. Type the organism name (or the brand name of your array) and click **`Next`**.

  c. Continue providing the information requested on each screen and click **`Next`** until you have completed the wizard. For details, see "Genome Wizard" on page C-1.

If you choose to skip this step, the Autoloader (used in Step 2) will load gene information directly from your data files. However, if you want to retrieve annotations for your genome using the GeneSpider (Step 4), you will have to enter the GenBank accession number of each gene in column 10 of the master gene table that was created by the Autoloader. Silicon Genetics can provide annotated genomes for many of the most commonly used arrays. Please call 1-866-SIG-SOFT or email support@sigenetics.com for details.

- **Step 2:** Load an Experiment

  a. Select **`File > Autoload Experiment`**.

  b. Choose a file.

  c. Either GeneSpring will recognize the format of your data file and ask you to name your genome, or you will have to set up columns using the column editor.

To Set Up Columns

1. Click each of the cells in Function row and choose a data type from the pull-down menu.
2. Click the **`Load Now`** button.

a. GeneSpring will ask you if you would like to load more files for this experiment. If you have additional files, click the appropriate box; otherwise click **`No, Load Only This File`**.

b. Enter an experiment name into the Choose Experiment Name window and click **`Save`**.

Copyright 1998-2001 Silicon Genetics

Alternatively, select **File > Manual Load Experiment > Experiment Import Wizard**. Follow the instructions on each screen until your experiment is loaded. For more information on using the Wizard, see "The Experiment Wizard" on page D-1.

- **Step 3:** Assigning Normalizations, Parameter Values, and Interpretations

  a. Select **Experiments > Experiment Normalizations**. Choose the types of normalizations to apply. Four classes of normalizations are available: background subtraction, per spot normalizations, per chip (global) normalizations, and per gene normalizations. Specify normalizations and save. For information about normalizations and when to apply them, see "Experiment Normalizations" on page 2-21.

  b. Select **Experiments > Change Experiment Parameters**. Set parameter units, values, value order, and add any missing parameters. For information about changing experiment parameters, see "Change Experiment Parameters" on page 2-8.

  c. Select **Experiments > Change Experiment Interpretation**. Select the mode of display, lower and upper bounds of data, the flagged measurements to be included, whether to use the Global Error Model, whether the data should be continuous, non-continuous, viewed as a replicate or color-coded. Note that these assignments are an extremely important preparation for any type of data analysis. For information about changing experiment interpretations, see "Changing the Experiment Interpretation" on page 2-17.

- **Step 4:** Annotate your genome (optional)

  Most researchers will want to import the maximum amount of biological information available about each gene before beginning analyses. After collecting the data, it is a good idea to make lists of genes based on appropriate keywords.

  a. Select **Annotations > GeneSpider**.

  b. Select a database from which to update your annotations. Then select the column in your master gene table that contains the accession number (usually Column 10 for the GenBank locus). Make sure there are accession numbers in the column you select.

  c. Click the **Start** button (the GeneSpider may continue gathering information for many hours). Remember to click **Save and close** when the GeneSpider is finished. For details on the GeneSpider see "Annotation Tools" on page 2-15.

At this point you are ready to begin working with your data.

Copyright 1998-2001 Silicon Genetics

Basic actions

Once you have loaded your data, GeneSpring will open a window with information from your new genome, and initially display all the genes in your experiment. If you just opened GeneSpring and want to see your new genome select **`File > Open Genome or Array`** and choose your genome from the pop-up list.

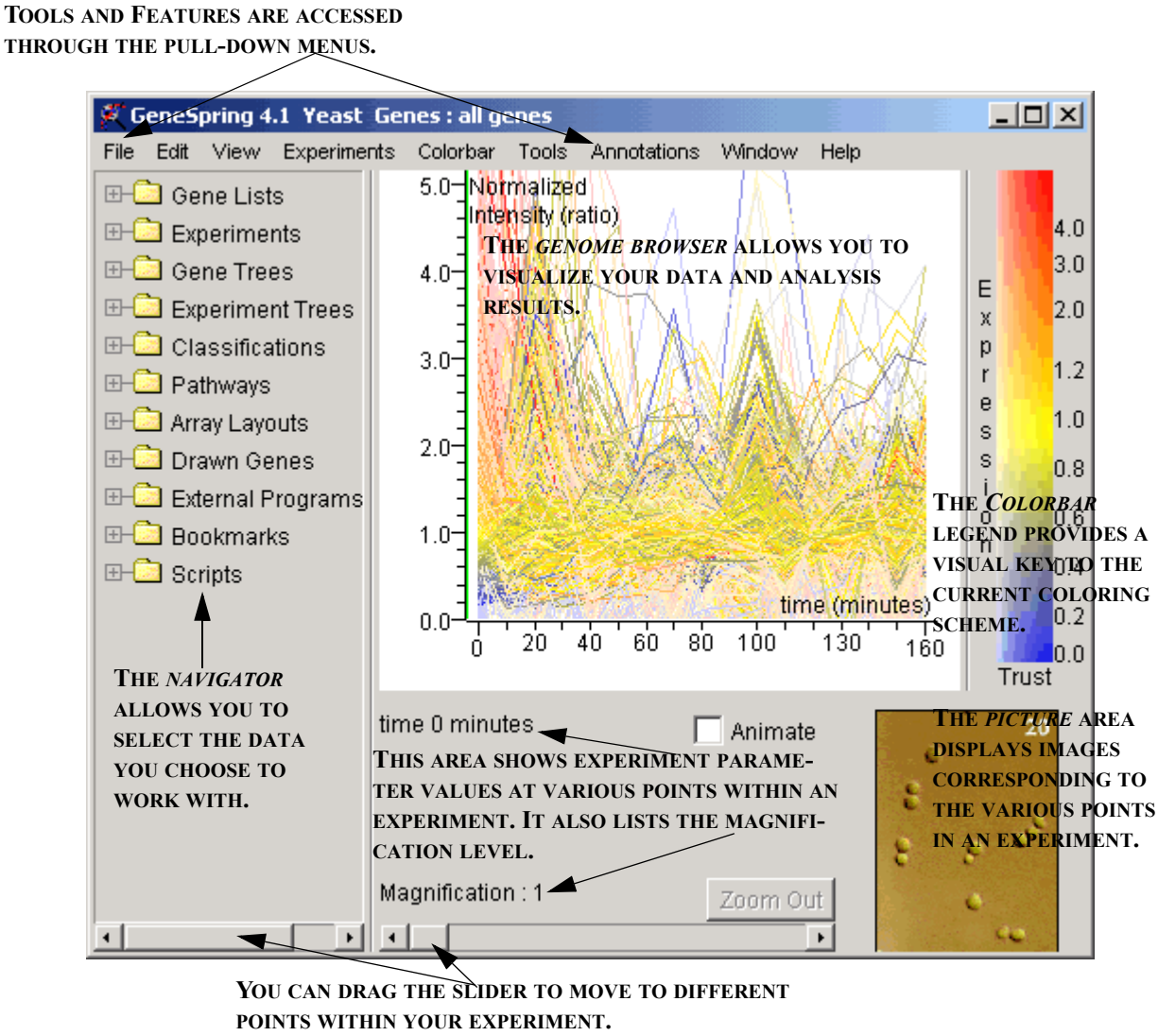

TOOLS AND FEATURES ARE ACCESSED THROUGH THE PULL-DOWN MENUS.
GeneSpring 4.1 Yeast Genes : all genes
File Edit View Experiments Colorbar Tools Annotations Window Help
Gene Lists
Experiments
Gene Trees
Experiment Trees
Classifications
Pathways
Array Layouts
Drawn Genes
External Programs
Bookmarks
Scripts
THE NAVIGATOR ALLOWS YOU TO SELECT THE DATA YOU CHOOSE TO WORK WITH.
Normalized Intensity (ratio)
THE GENOME BROWSER ALLOWS YOU TO VISUALIZE YOUR DATA AND ANALYSIS RESULTS.
time (minutes)
Expression
Trust
THE COLORBAR LEGEND PROVIDES A VISUAL KEY TO THE CURRENT COLORING SCHEME.
time 0 minutes
Animate
THIS AREA SHOWS EXPERIMENT PARAMETER VALUES AT VARIOUS POINTS WITHIN AN EXPERIMENT. IT ALSO LISTS THE MAGNIFICATION LEVEL.
Magnification : 1
Zoom Out
THE PICTURE AREA DISPLAYS IMAGES CORRESPONDING TO THE VARIOUS POINTS IN AN EXPERIMENT.
YOU CAN DRAG THE SLIDER TO MOVE TO DIFFERENT POINTS WITHIN YOUR EXPERIMENT.

Figure 1-3 The main GeneSpring window

Below are some basics to get you moving around GeneSpring.

- **Changing the genes displayed:** Open the gene list folder in the navigator. GeneSpring initially displays the “all genes” list. You can change the genes shown in the display by choosing another list.
- **Views:** You can change the view in the genome browser using the View menu. GeneSpring initially displays the Classification view, where genes are displayed according to pre-defined categories. However, you can view displayed genes as a graph, a scatter plot, a bar graph, an

ordered list, etc. Note that some views such as Tree, Pathway, and Array Layout require some preparation, such as creating a tree or adding a pathway or Array Layout image. For details on views, see "Viewing Data in GeneSpring" on page 3-1.

- **Zooming in:** To zoom in on a region or gene, click on an area and drag your cursor diagonally. You will see an expanding rectangle. Release the mouse and GeneSpring will zoom in on the region enclosed by this rectangle.
- **Zooming out:** To zoom out, right-click (Control + click for Mac) and choose **`Zoom Out`** to go back one level or **`Zoom Fully Out`** to zoom out as far as possible.
- **Moving around the screen:** You can move around a zoomed-in screen by using Page Up, Page Down and the arrows keys.
- **Selecting a gene:** Click once on a single gene to select it.
- **Selecting multiple genes:** Hold down the Shift key and drag to select multiple genes. Or hold down the Shift key and click on individual genes to select them one by one.
- **Finding a specific gene:** Select **`Edit > Find Gene`**. Type in the gene name or keyword and click **`OK`**. GeneSpring will select and zoom in on the gene.
- **Inspecting genes:** You can view detailed information about a gene by double-clicking on it and bringing up the Gene Inspector window. This is easier after zooming in on the gene. A shortcut to the Gene Inspector is Ctrl + I, or ⌘+I for Mac users.
- **Undo:** You can undo your last action by selecting **`Edit > Undo`** or **`Ctrl + Z`** (⌘ + Z for Mac users).

Your First Gene Lists

To make lists from appropriate keywords:

1. Select **`Annotations > Make Gene Lists from Properties`**.
2. Choose the property you would like to use for generating lists and click **`OK`**.

To make a list based on biological function:

1. Select **`Annotations > Build Simplified Ontology`**.
2. Name your new list and click **`OK`**.

To make lists from a group of selected genes:

1. While the group of genes is still highlighted, right-click over the highlighted area and select **`Make List from Selected Genes`** from the pop-up menu.

You will find your new lists in the Gene Lists folder.

Copyright 1998-2001 Silicon Genetics

Tips for Mac Users

Except where otherwise noted, instructions in this manual describe GeneSpring usage on a PC. If you are a Mac user, you will find the following keystroke and mouse conversion information helpful:

- **Right-Click:** Hold the Control button and click. This will most often activate a pop-up menu.
- **Ctrl = ⌘ :** Wherever the manual mentions Ctrl, for example press Ctrl + I to reach the Gene Inspector, substitute the ⌘ (Apple) key for Ctrl.
- **Drawing genes on a pathway:** Hold down the Option key and drag your cursor diagonally to draw a gene on a pathway. See “Pathways” on page 4-23 for more information.

Note that on a Macintosh computer the menu bar is at the top of the screen, not on the individual GeneSpring windows as displayed in this manual.

The Navigator

GeneSpring organizes data elements relating to your genome into folders in the navigator. Each folder contains a specific type of information. The labeled diagram and list below briefly explains the purpose of each folder.

Copyright 1998-2001 Silicon Genetics

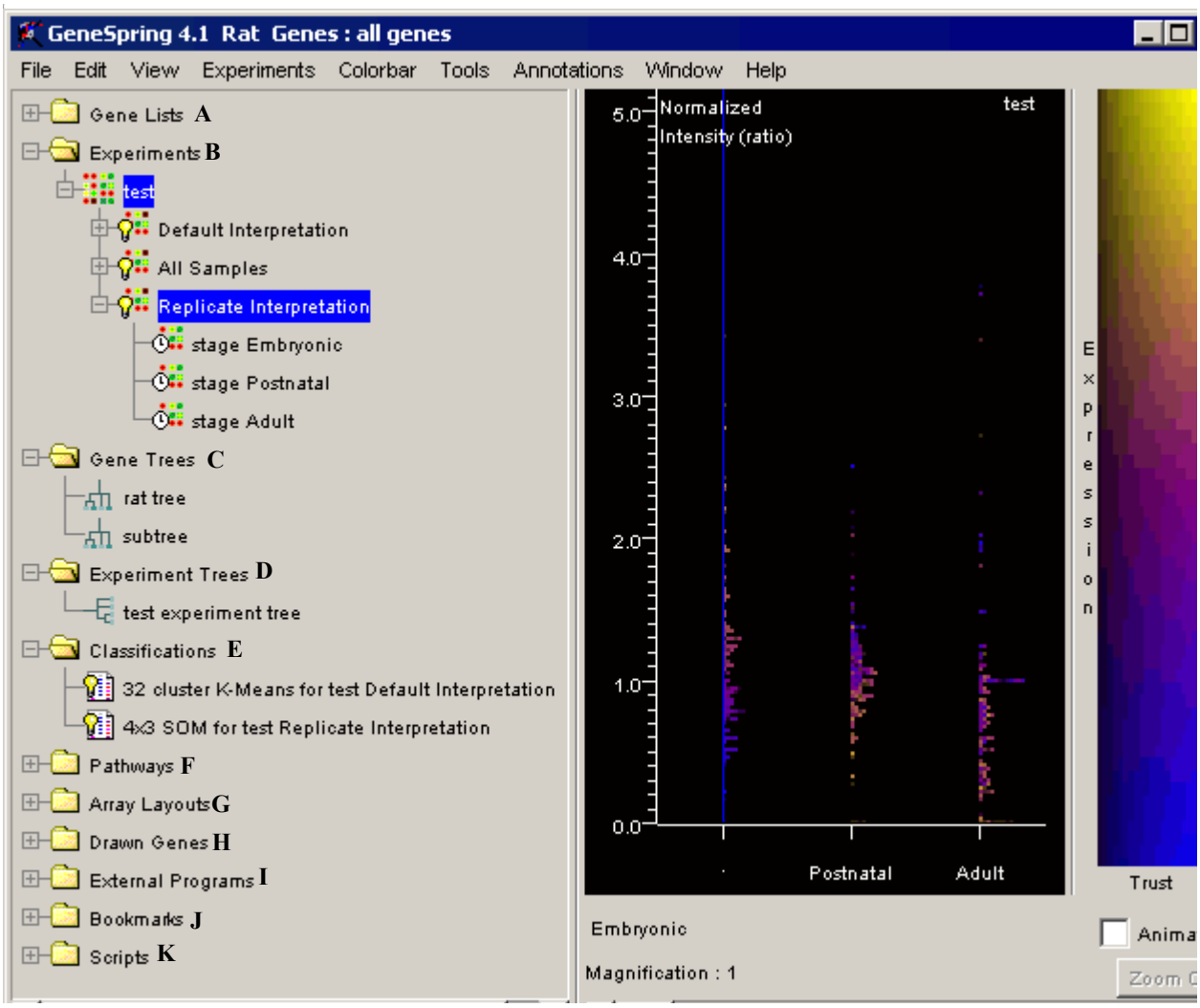


Figure 1-4 The GeneSpring Navigator

A. During analysis, you will create and work with interesting collections of genes known as gene lists. These gene lists are stored in the Gene Lists folder. By default, GeneSpring makes and displays an "all genes" list containing all genes in the genome.

B. The Experiments folder contains experiment information. Experiments are divided into *interpretations*. Experiment Interpretations tell GeneSpring how to treat and display your experiment variables, called experiment *parameters*.

Conditions are groupings of one or more samples. Each sample may be a condition, as in the "All Samples" interpretation or a condition may include multiple samples. For example, because the experiment above is organized according to the parameter values Embryonic, Postnatal and Adult, these can be called the *conditions* the experiment. Within these conditions, the parameter *day* is being treated as a replicate and has been averaged for each condition, Embryonic, Postnatal and Adult, across all samples. Hence a condition can include data from more than one sample.

Copyright 1998-2001 Silicon Genetics

C. Any gene trees created in GeneSpring are kept in the Gene Trees folder. Gene trees are dendrograms used as a method of showing relationships between the expression levels of genes over a series of conditions.

D. Experiment trees are like gene trees, except that instead of showing the relationships between genes, they show the relationships between the expression levels of samples. Experiment trees are kept in the Experiment Trees folder.

E. The Classifications folder contains genes that have been grouped or *classified* to divisions defined by k-means or SOM clustering.

F. Pathways are images of regulatory or metabolic pathways that can be imported into GeneSpring. Genes are overlaid on these images allowing you to observe their changing expression levels across experimental conditions. A feature called **`Find Genes Which Could Fit Here`** can be used as a tool to predict new pathway elements.

G. The Array Layouts folder contains information about the arrangement of the spots on your array. These can be used to recreate an image of your arrays to check for regional abnormalities.

H. Drawn genes are lines representing gene profiles that you draw in the genome browser. You can then search for genes matching that profile. Any drawn genes you create are stored in the Drawn Genes folder.

I. External programs are analysis programs outside GeneSpring that can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

J. Bookmarks are saved display settings such as experiment, gene list, color scheme, selected genes, etc. You can always save your current display and return to it later by opening the Bookmarks folder and selecting a particular bookmark.

K. Scripts are tools that save time by allowing a long series of data analysis steps to be performed at once. Scripts are re-usable and can be applied to any data set. You can create your own scripts using Silicon Genetics Script Editor. All scripts, including complimentary scripts shipped with GeneSpring 4.1, are stored in the Scripts Folder.

By default, folders in the navigator are closed, although on start-up GeneSpring displays an “all genes” or “all genomic elements” gene list. You can change the default genome that GeneSpring initially opens by going to **`Edit > Preferences`**, selecting **`Data Files`** from the pull-down menu, and typing a genome name in the Default Genome text field.

The GeneSpring Hierarchy of Objects or, Where Is My Data Stored?

Understanding the GeneSpring file structure can be helpful for installing, updating and working with GeneSpring. In your Programs folder (Windows) or Applications folder (Mac OS), you will find the Silicon Genetics directory, containing GeneSpring and jre.

Copyright 1998-2001 Silicon Genetics

The GeneSpring folder contains bin, data, docs and UninstallerData folders. The principal GeneSpring program file (*GeneSpring.jar*) is kept in the bin folder. License keys belong in the data folder and documentation is stored in the docs folder.

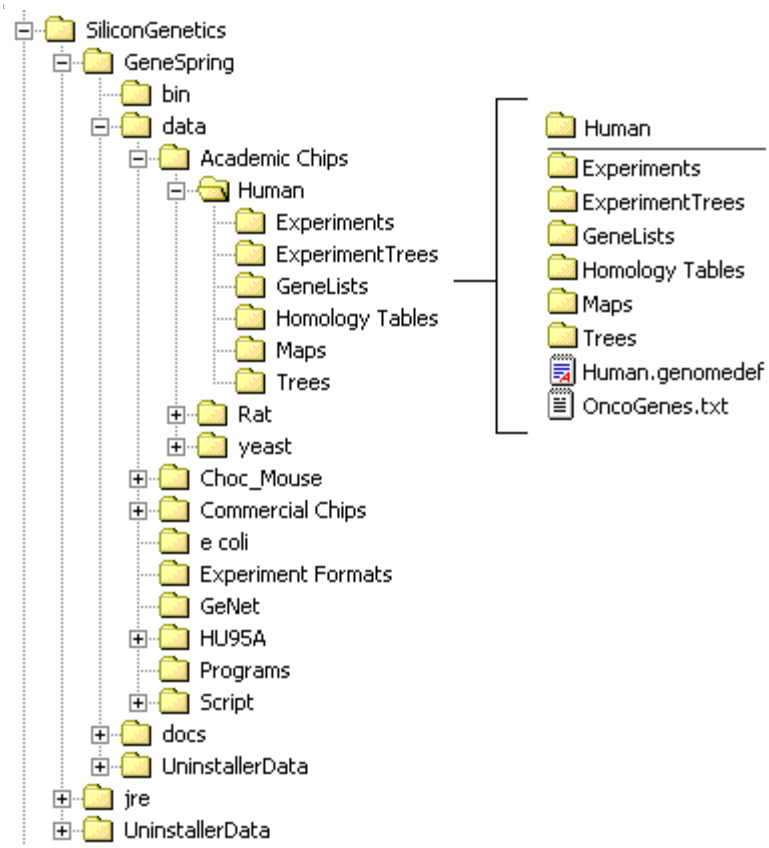


Figure 1-5 GeneSpring's internal data structure

The data folder is also important because this is where all the information about your genomes and experiments is stored. Each genome or organism folder contains two key files: the genome definition file (*.genomedef*) and the master table of genes (*.txt*), along with folders containing information relating to experiments, maps, trees, gene lists, and other data relevant to the particular organism.

Copyright 1998-2001 Silicon Genetics

Commonly Used GeneSpring Functions

To open a different genome, choose **File > New Genome**. To open another copy of the main window, choose **File > New Linked Window**. Each of these will bring up a new main window similar to the one described in "GeneSpring Basics" on page 1-7.

To change preferences (colors, start up genome, etc.), choose **Edit > Preferences**. See Appendix B, "Preferences Window" for more details.

The Gene Inspector window

Double-clicking a gene will bring up the Gene Inspector window. This window contains specific information about the selected gene. See "Gene Inspector" on page 3-37 for details. Information presented in the Gene Inspector might include:

- knowledge you have about your selected gene (typically text).
- graphs of the selected gene's expression profile from the current experiment.
- links to internet or intranet databases on the web for the selected gene.

Making Lists

There are many ways to create a list of genes, see Chapter 4, "Analyzing Data in GeneSpring" for more details. From the Gene Inspector window you can do the following.

- **Making Lists with the Find Similar Command:** The **Find Similar** command allows you to create a list of genes having similar expression profiles to the gene being displayed. See "Making Lists with the Find Similar Command" on page 4-13 for more details.
- **Making Lists with the Complex Correlation Command:** The Complex Correlation Command allows you to make a list of all the genes satisfying various conditions you define. See "Making Lists with the Complex Correlation Command" on page 4-14 for more details.

Many other tools are available with which you can make lists.

- **Making Lists with the Venn Diagram**—Select **Colorbar > Color by Venn Diagram** to begin. Right-clicking over lists in the navigator will allow you to fill the diagram. This function allows you to make lists based on the membership of genes in a Venn Diagram. See "Making Lists with the Venn Diagram" on page 4-19 for more details.
- **Making Lists with the Filter Genes Command**—Select **Tools > Filter Genes**. It allows you to use expression level constraints and control strength restrictions to create a smaller gene list. See "Filter Genes Analysis Tools" on page 4-1 for more details.
- **Making Lists from Selected Genes**—You can make a list of all the genes you have selected in the genome browser by right-clicking and choosing **Make List from Selected Genes**. See the "Finding and Selecting Genes" on page 3-4 for how to select genes. See "Making Lists from Selected Genes" on page 4-22 for more details on this method of making a gene list.
- **Making Lists from Conjectured Regulatory Sequences**—Once you have found possible regulatory sequences using the *Find Potential Regulatory Sequences* window (see "Regula-

Copyright 1998-2001 Silicon Genetics

tory Sequences" on page 4-26 for more details) and are inspecting one of the sequences in the *Conjectured Regulatory Sequence* window, you can make a list of all of the genes containing that sequence by selecting **`List > Make Gene List`**. See "Using the Conjectured Regulatory Sequence window" on page 4-29 for more information.

Copyright 1998-2001 Silicon Genetics

Chapter 2 Creating DataObjects in GeneSpring

The Experiment Autoloader

The Experiment Autoloader is a time-saving feature that is programmed to automatically recognize and load most data formats. The Autoloader automatically recognizes the following formats:

- Clontech AtlasImage 2.0
- Affymetrix Metrics
- Affymetrix Pivot
- Axon GenePix 4000
- BioDiscovery Imagene 4
- Incyte Internet
- Incyte GEM Tools 2.4
- Packard Biochip (GSI Lumonics) ScanArray
- Packard Biochip QuantArray 4000
- Generic one-color
- Generic two-color

If the Autoloader is unfamiliar with your file format, you can use the Column Editor to specify the type of data in each column. Once the Column Editor learns the location and identity of the relevant columns of data, it adds these specifications to its list of known file types so that you can load subsequent experiments in batch.

Make sure you use the raw, tab-delimited files just as they come out of the scanner, as GeneSpring uses the information in the column headers. If you have cut out header information, you will need to find your original tab-delimited data files and use those.

To Autoload an Experiment

1. Select **File > Autoload Experiment** or **Ctrl+O**.
2. 2. Choose the data file or folder you wish to load. Make sure all the files have exactly the same format.
3. If GeneSpring correctly identifies your file format, click **Yes**. The Select Genome window will appear.
   - If GeneSpring does not correctly identify your file format, choose **No**. A dialog box will appear asking you to set up column formats for your data or use the Experiment Import Wizard.
     a. If all your files are in the same format, choose **Yes**. This will bring up the Column Editor. See “To set up Column Formats” on page 2-2.

Copyright 1998-2001 Silicon Genetics

b. If your files are not in the same format, choose **No**. This will exit the Autoloader. You will need to use the Experiment Import Wizard, “The Experiment Wizard” on page D-1 for details.Choose an existing genome or create a new one.

4. Choose an experiment name and click **Save**. Your experiment will appear in the genome browser.

To set up Column Formats

If GeneSpring does not recognize your file format, you can use the Column Editor to assign headings and functions to each column in your data file.

The Column Editor is programmed to remember the format of your file for the next time you load data with that format. Note, however, that the Column Editor will not remember a format if you have more than one sample in a file or if you have more than one signal column.

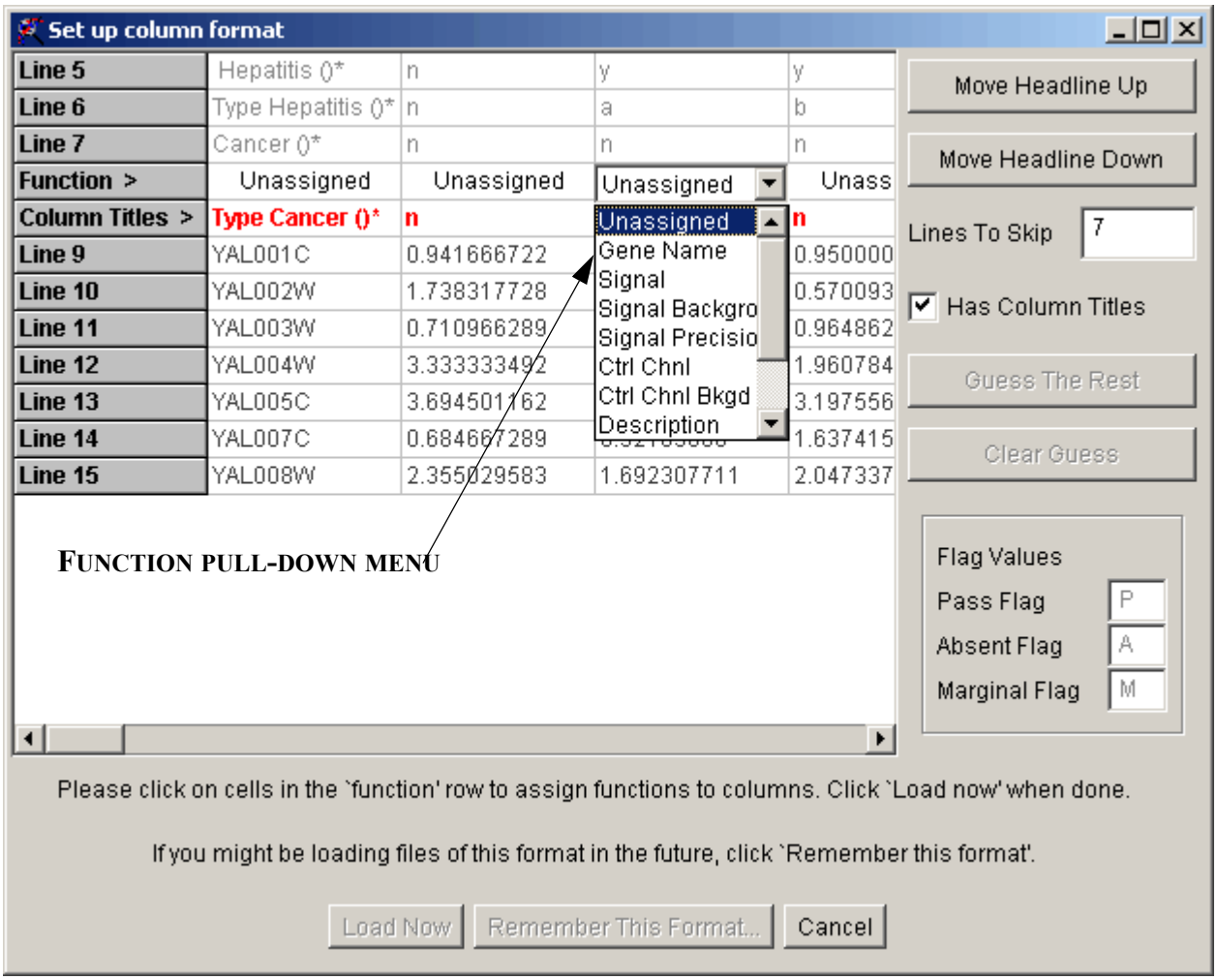


Figure 2-1 The Column Editor

GeneSpring will have guessed which row represents your column titles. If GeneSpring is incorrect, click the **Column Titles** cell at the far left. Use the **Move Headline Up** or **Move**

 Copyright 1998-2001 Silicon Genetics

Headline Down buttons to select a new row to use as column titles. If your file has no column titles, deselect the check box marked Has column titles.

1. In the row marked Function, you can assign functions to each column. Choose a function from the pull-down menu in each column. (See Figure 2-1.) You can have unlimited Flag and Unassigned columns, however other functions can only be used once. At least one Gene Name column and one Signal (raw data) column are required.
    - If you assign a Flag column, you will be able to specify the letter or number indicating Present, Absent and Marginal calls.
2. After your initial assignments, click the **Guess the Rest** button and GeneSpring will attempt to label the remaining columns. If GeneSpring is incorrect, click the **Clear Guess** button to remove the column labels.
3. If you wish to use the same format in the future, select **Remember This Format**. This format will be added to the cache of recognized formats and GeneSpring will suggest it in the future. Note, however, that the Column Editor will not remember a format if you have more than one sample in a file or if you have more than one signal column. GeneSpring will ask you to name your format.
4. Click **Load Now** to load the experiment. The Select Genome window will appear.
5. Choose an existing genome or create a new one.
6. Choose an experiment name and click **Save**. Your experiment will appear in the genome browser.

After loading an experiment, examine and change your normalizations, interpretations, and parameters.

- To change normalizations, select **Experiments > Experiment Normalizations**. See “Experiment Normalizations” on page 2-21 for details.
- To change parameters, select **Experiments > Change Experiment Parameters**. See “Change Experiment Parameters” on page 2-8 for details.
- To change interpretations, select **Experiments > Change Experiment Interpretation**. See “Changing the Experiment Interpretation” on page 2-17 for details.

Autoloader Normalizations

The Autoloader will normalize your new files based on the technology used to create the original data files.For more information on normalizations, see “Experiment Normalizations” on page 2-21.

 Copyright 1998-2001 Silicon Genetics

One-Color Experiments

One-Color normalizations will automatically display all information flagged as Present or Unknown:

- **Per-chip:** Distribution of all genes using 50th percentile,cutoff = 10
- **Options:** Use background correction if necessary, anything but absent
- **Per-gene:** Median for each gene, cutoff = 0.01 (if 2+ samples).

Two-Color Experiments

Two-color experiments are automatically normalized to a signal ratio. Two-color normalizations will automatically display all information flagged as Present or Unknown:

- Per-spot: Use control channel to calculate ratio, cutoff = 10
- Per-chip: Distribution of all genes using 50th percentile,cutoff = 0.01
- Options: Use background correction if necessary, anything but absent

Default Normalizations of Commercially Available Products

Affymetrix

Pivot Table will automatically display all information flagged as Present or Unknown:

- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 10
- **Options:** Use background correction if necessary, anything but absent
- **Per-gene:** Median for each gene, cutoff = 0.01 (if 2+ samples). By default, GeneSpring forces negative values to zero.

Metrics will automatically display all information flagged as Present or Unknown:

- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 10
- **Options:** Use background correction if necessary, anything but absent
- **Per-gene:** Median for each gene, cutoff = 0.01 (if 2+ samples). By default, GeneSpring forces negative values to zero.

Axon

GenePix 4000 will automatically display all information flagged as Present or Unknown:

- **Per-spot:** Use control channel to calculate ratio, cutof f= 10
- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 0.01
- **Options:** Use background correction if necessary, anything but absent

Copyright 1998-2001 Silicon Genetics

BioDiscovery

Imagene 4 will automatically display all information flagged as Present or Unknown:

- **Per-spot:** Use control channel to calculate ratio, cutoff = 10
- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 0.01
- **Options:** Use background correction if necessary, anything but absent

Incyte

GEMTools 2.4 will automatically display all information flagged as Present or Unknown:

- **Per-spot:** Use control channel to calculate ratio, cutoff = 10
- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 0.01
- **Options:** Use background correction if necessary, anything but absent

Internet Download will automatically display all information flagged as Present or Unknown:

- **Per-spot:** Use control channel to calculate ratio, cutoff = 10
- **Per-chip:** Distribution of all genes using 50th percentile, cutoff = 0.01
- **Options:** Use background correction if necessary, anything but absent

Replicates

If you have three or more experiments with the same samples, GeneSpring will automatically normalize to the median for each gene. Please refer to“Dealing with Repeated Measurements” on page G-16 for a mathematical explanation of this process.

Remembered Formats

While you cannot edit remembered formats, you can share them. (If you need to change a remembered format, you will have to build a new one.) To share remembered format files, use your favorite browser or file management program to copy the file from:

YourLocalDrive:\Program Files\SiliconGenetics\GeneSpring\data\Experiment Formats\name.expformat

You can then paste the file into a shared drive.

Copyright 1998-2001 Silicon Genetics

Merging, Splitting and Duplicating Experiments

The Merge/Split Experiments function allows you to merge or split experiments or groups of experiments in their entirety or by condition. Note that only conditions from your default interpretation are available for merging/splitting. GeneSpring also allows you to duplicate experiments.

Once you merge an experiment you can treat it like any other experiment with a few notable exceptions. If you have multiple spots for one gene on a single chip, GeneSpring will only retain the median of those values in the merged experiment. This means that you will not have access to error bars. Also, GeneSpring will only be able to access data from the following columns: gene name, signal, signal background, signal precision, control channel, control channel background, description, GenBank ID, flags, and region.

To Merge or Split an Experiment

1. Select **Experiments > Merge/Split Experiments**.
2. To merge experiments/conditions, open the Experiments folder in the mini-navigator and click on the first experiment folder, experiment or condition you would like to merge (find a condition by clicking on the plus sign next to the experiment icon).
   - Click the **Add** button.
   - Repeat steps 3 and 4 below until you have added all your experiments/conditions.

To Split Experiments/Conditions

1. Select **Experiments > Merge/Split Experiments**.
2. Open the Experiments folder in the mini navigator and click on the first experiment/condition you would like to delete.
   - Click the **Add** button.
   - Individually select the experiments/conditions you would like to remove and click the **Remove** button.
3. Click **OK**. The Experiment Parameters window will appear. You will see a parameter called *Experiment* listing the names of the experiments involved. You can alter, add, or delete parameters. For information about the functions in this window, see “Change Experiment Parameters” on page 2-8.
4. Click **Save**. The Choose Experiment Name window will appear.
5. Enter names for your experiment and experiment folder and click **Save**. You will find your merged/split experiments in your Experiments folder.

 Copyright 1998-2001 Silicon Genetics

To Duplicate an Experiment

1. Select **Experiments > Duplicate Experiment**. Right-click the experiment name and select **Duplicate Experiment** from the resulting pop-up menu. The Duplicate Experiment dialog box will appear.
2. Name your experiment or accept the default.
3. Click **OK**. Your new experiment will appear in the Experiments folder in the navigator.

Loading from Subchips

Sometimes, due to oddities in the way region normalizations are done, you will need to enter each chip as a separate experiment and merge them together.

Creating a Genome through the Autoloader

In GeneSpring, a genome includes all the genes on your chip. When you create a genome through the Autoloader, GeneSpring creates a genome on the fly based on genes in your experiment data files. This means that unlike a genome created in the New Genome Installation Wizard, a genome created through the Autoloader has no annotations and no means of obtaining annotations from public databases. The genome consists of a master table of genes and a genome definition file. If you create a genome through the Autoloader after accepting a file format recognized by GeneSpring, anything not standard to that recognized format will not be included in the master table of genes. (The master table of genes contains all the information associated with genes in a given genome.) For example, if GeneSpring recognizes an Affymetrix file, but that file has GenBank accession numbers, the numbers will not be loaded. You can add these numbers later to column 10 of the master table of genes. (If your data files have a description column, the Autoloader will include it in the master gene table.)

If you have difficulties creating a genome through the Autoloader, you can use the New Genome Installation Wizard, see “Genome Wizard” on page C-1.

To Create a Genome Through the Autoloader

Start the autoloader:

1. Select **File > Autoload Experiment**.
2. Choose the data file you wish to load.
3. Verify the file format. For details, see “The Experiment Autoloader” on page 2-1.

Create your genome:

4. Select a genome from the Select Genome window in the autoloader. If your genome is not listed, enter the new genome name. Click **Choose Selected Genome**.
    - If you have entered a new genome, a second window will ask if you want to continue. Click **Yes**.

Copyright 1998-2001 Silicon Genetics

5. You will have an option to load additional files. Choose the files you wish to load. GeneSpring will add genes in these data files to the genome.

Change Experiment Parameters

You will want to use the Change Experiment Parameters window to assign parameter names and units (e.g., time and minutes) to your data. (For an explanation of parameters in GeneSpring, see "Definitions of Parameters" on page 2-11.) You can also use this window to add and delete parameters and rearrange the order of non-numeric parameter values on the horizontal axis. The Change Parameters window has an Edit menu with a variety of options including the Extract Sub-values feature, which can conveniently automate your parameter assigning process if you set up your file names as described below.

To Change Experiment Parameters

1. Select **`Experiment > Change Experiment Parameters`**.
2. Fill in the Parameter Name and Parameter Units (the latter only if applicable).
3. In the Numeric and Logarithmic rows, select **`Yes`** or **`No`** from the drop-down menus. You can also paste data in the **`Sample`** cells.
4. Click **`Save`** to change the parameters in your current experiment or **`Save As`** to save this parameter set-up as a new experiment.

To add a parameter, click the **`Add Parameter`** button.

To delete a parameter, click the gray bar above the column you would like to delete and then click Delete Parameter.

To rearrange the order of non-numeric parameters on the horizontal axis, click **`Set Value Order`**. To Sort Ascending/Descending, first click the gray bar at the top of the column. To move individual entries, click on the entry then select one of the move buttons: Move Up, Move Down, Move To Top, or Move To Bottom.

You also have several options under the Edit menu at the top of the window:

- **Cut:** Allows you to delete entries one at a time or as a group (to do the latter, click on one entry and then hold the Shift key down while clicking on additional entries).
- **Copy:** Allows you to copy an entry for pasting in another cell.
- **Paste:** Allows you to paste a previously copied entry.
- **Paste Transposed:** Allows you to copy a row from a tab-delimited text file or spreadsheet and paste it into a column.
- **Clear:** Clears selected cell.
- **Replace:** Allows you to replace many entries at once. Select the entries you wish to change and choose **`Replace.`**

Copyright 1998-2001 Silicon Genetics

Or, to replace all instances of an entry, choose **Replace** and then deselect the **Replace in selected cells only** checkbox before clicking **OK**.

- **Extract Sub-values:** This feature automates parameter assignment. To use it you must create file names based on your parameter values (e.g., Rlr001a.txt, where "Rlr0" refer to an experiment and "01" is your sample number and "a" is the region designator).

When you implement the Extract Sub-values feature, file names are broken down into sub-values. GeneSpring is programmed to first look for alternating constant fields and variable fields and to make parameters out of the variable fields. Next it divides the variable fields into groups consisting of uninterrupted stretches of either numbers, letters, or non-alpha-numeric characters and makes parameters out of each of these groups.

- **Fill Down:** Allows you to replace entries using the top selected cell. Click on the cell you would like to use as the replacement and then, holding down the **Shift** key, click on the cells underneath whose values you would like replaced with the original cell.
- **Fill Sequence Down:** Allows you to fill down as described above, but additionally will recognize a simple numeric or alphabetic sequence and continue it.

The Experiment Parameters Window

To reach the Experiment Parameters window, select **Experiment > Change Experiments Parameters**.

There are four special rows at the top of the Experiment Parameters window.

- **Parameter Name:** This box should be filled with a short description of the parameter. It will be used in the main GeneSpring navigator, it will be much easier to read later if you use short names or names with distinctive beginnings. You can paste or type directly in this text box.
- **Parameter Units:** These are any units that will apply to the parameter values. For example, the parameter values of drug concentration could be 10 ppm, 20 ppm, 30 ppm and 40 ppm. You can paste or type directly in this text box.
- **Numeric:** Selecting this cell will result in a yes/no drop-down menu. Choose one or the other the indicate whether or not the parameter values are numeric. If you click **Yes**, GeneSpring will automatically order the parameter values in numeric order from smallest to largest. Please refer to "Re-order the Parameters" on page 2-10 before you make an permanent decisions.
- **Logarithmic:** Selecting this cell will result in a yes/no drop-down menu. Choose one or the other the indicate whether or not these parameter values should be displayed on a logarithmic scale.

Copyright 1998-2001 Silicon Genetics

Add a Parameter

Click the **`Add Parameter`** button at the bottom of the window and a new column will appear at the far left.

You can paste in columns of information by clicking the cells of the *Sample* section. For example, if you had an Excel spreadsheet of data and wanted to copy and paste a column from it, you could copy a large section of column and paste it into the new column. You can also copy information out. You can only add columns (parameters and parameter values), you cannot add rows (samples) into this table.

Re-order the Parameters

To change the order of your parameters as they are displayed in along the X-axis in the main GeneSpring window, you will need to select an entire column or part of a column and then use the **`Set Value Order`** button at the very bottom of this panel.

Sort Descending

For example, if you wanted to show the numeric, continuous parameter “Kryptonite Concentration” in reverse order (40, 30, 20, 10, 0) of the normal arrangement (0, 10, 20, 30, 40) you first need to change the setting to a non-numeric parameter and select the column by clicking on the gray bell at the very top. You cannot change the order of a parameter defined as numeric.

To select part of a column you can highlight it in the normal fashion, or while holding down the **`Shift`** key click in the top most cell you want. GeneSpring will *select down* the column for you. Click the **Set Value Order** button.

Select all the values you want to order so you can use the *Sort Ascending* or *Sort Descending* buttons. The main GeneSpring window will sort your parameters according to the new system.

Sorting Manually

You may select just one of the parameter values in the main window of the Parameter Value Order box and use the move up/move down buttons to arrange the order to your liking.

Copyright 1998-2001 Silicon Genetics

Definitions of Parameters

Parameters are the variables you use to describe your experiment.

Parameter Vocabulary

- **Experiment parameters:** variables that can incorporate many sample parameter variables. Generally speaking, when the term parameter is used, it means an experimental parameter. As an example, parameters could be:
  - Kryptonite Concentration
  - Variety of Yeast
  - Andromeda Strain Infection
  - Test Repeat Number
- **Parameter-value:** is one of the possible values assigned to a variable. As an example, the parameters-values from the previous list could be:
  - Kryptonite Concentration in ppm, 0, 10, 20, 30, 40
  - Variety of Yeast, A or B
  - Andromeda Strain Infection, Healthy or Infected
  - Test Repeat Number, 1 or 2
- **Sample parameters:** variables used to describe the precise condition under which each sample (or measurement) was taken. You may have many parameter values applying to a single sample (such as time, drug concentration, etc.).

  The sample parameters are listed in the main GeneSpring navigator for every condition. Please refer to “Parameter Display Options” on page 2-12 for more details.

Parameters Displayed in the Navigator

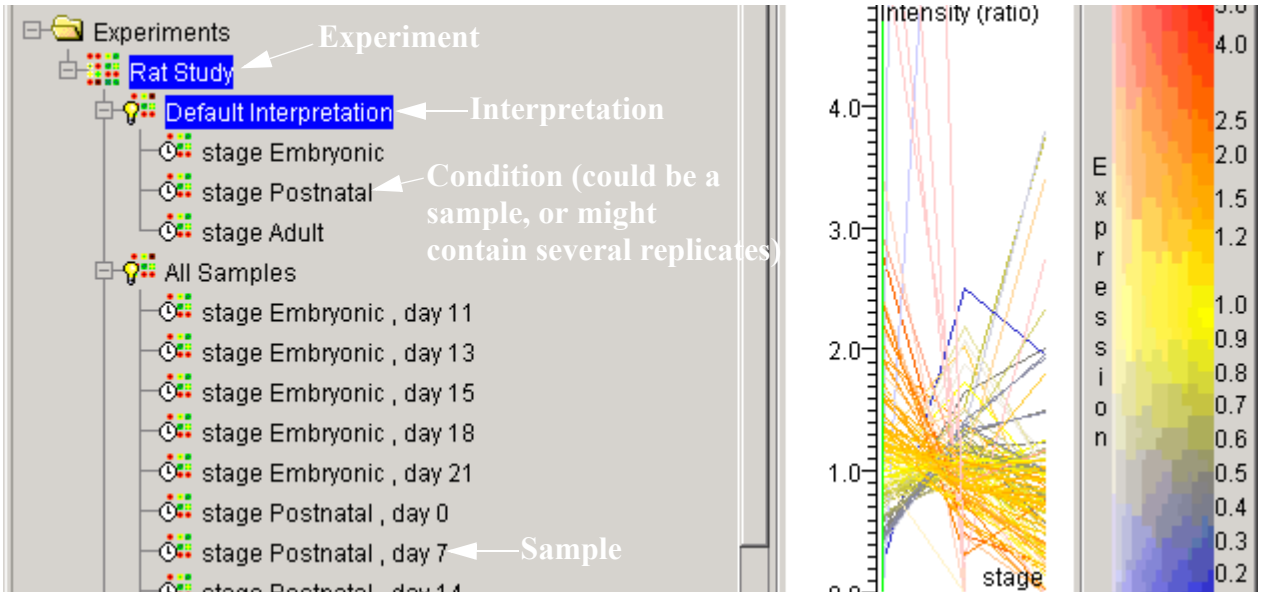


Figure 2-2 Data objects in the navigator

Copyright 1998-2001 Silicon Genetics

- **Measurement:** The smallest unit of data used by GeneSpring, you will only see measurements as the raw values present in the upper right table in the Gene Inspector. In the Graph view this will be presented as one point on one gene's line. (It may be easier to think of this as one spot or set of probes on one array.) A measurement is a number, such as 7.3.
  If you have no replicates, 1 measurement = 1 raw value = 1 spot on a chip.
- **Array:** a set of spots on a chip, typically expressed as a set of intensity measurements. An array typically has one sample on it. If you have gross slide problems, please see **"Array Layout View" on page 3-22** for more information. If all of the interesting genes of the genome fit onto one array, then the terms array, chip and sample can be considered synonymous.
- **Sample:** The data generated from a biological object placed onto an array or set of arrays. A sample's data is visible in the GeneSpring navigator, under the All Samples icon.
- **Condition:** A unique combination of parameters as applied to your sample. Each condition may be a single sample or a group of replicate samples combined based upon the parameter values defined for each sample. The easiest way to think of this is as the parameters under which the sample(s) was observed. If you have no replicates, condition and sample can be considered synonymous. In Figure 2-2 the conditions are Embryonic, Postnatal and Adult.
- **Interpretation:** A description of how GeneSpring displays the data for you to view. It would include a definition of applicable parameters and how the normalized numbers should be treated. This is the way a set of conditions is grouped. In Figure 2-2 the interpretation is the Default Interpretation.
- **Experiment:** a set of samples, generally designed to answer specific types of questions. The data are usually (but not always) manipulated in a normalized form. In Figure 2-2, the experiment is the Rat Study.

A Note on Multiple Parameters

The more experimental parameters you have, the more options you have for visually querying your data. If you have samples of tissues infected with the different disease possibilities such as (breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, syphilis, and no disease) you might want to use several experimental parameters for this experiment. Using multiple parameters (even if they all refer to the same information) allows you to group the data in many different ways which may give you different insights into your data set.

Parameter Display Options

GeneSpring offers four ways of visually displaying a parameter: a continuous element, a non-continuous element, a replicate (or hidden) element, or a color code. When you enter a new experiment in the Experiment Wizard, you will be asked which display option is most appropriate for each of your parameters. Your chosen display option will become the default display for that parameter. If you simply paste in a new experiment, all the parameters will be assigned the continuous display option. Regardless of how a parameter is entered in GeneSpring, you can change how each parameter is displayed within GeneSpring using the **`Experiment > Change Experiment Interpretation`** command. For more details on this, see "Changing the Experiment Interpretation" on page 2-17.

Copyright 1998-2001 Silicon Genetics

Replicate or Hidden Element

Parameters defined as replicated are averaged together and appear as a single parameter. A parameter defined as a replicate is graphically a hidden variable. Defining a parameter as a replicate is the easiest way to deal with repeated samples inside GeneSpring.

The equation used for averaging repeated samples is exactly the same one used to average repeated measurements in a raw data file. See "Dealing with Repeated Measurements" on page G-16 for more information. The only difference is the averaging done to repeated parameters is done *after* the raw data has been normalized.

Continuous Element

A continuous variable is one where each value of the experimental parameter exists in series on a continuum with the other values in that experimental parameter, rather than as discrete points. Each parameter-value is related to the parameter values on either side of it and adjacent data points are connected together by lines. Typically, continuous variables are numeric. This requires the parameter values be in a particular order. GeneSpring will automatically order numerical parameters from highest to lowest, and order non-numerical parameters in alphabetical order. When graphing by a continuous parameter each parameter-value is placed on the X-axis, in order, from left to right. You can change this default order, please refer to "Re-order the Parameters" on page 2-10 for more details.

Non-Continuous Element (Set)

A non-continuous (or set) variable is when each parameter-value of the experimental parameter exists independent of each other, as discrete points. When a non-continuous element is graphed, each parameter-value is placed on the horizontal-axis, in order, from left to right. GeneSpring will automatically order numerical parameters from highest to lowest, and order non-numerical parameters in alphabetical order. See "Re-order the Parameters" on page 2-10 if you wish non-numerical parameter values to be graphed in a particular non-alphabetical order.

When displaying data from a non-continuous parameter, data points are graphed in histograms, as discrete points. A gene deletion is a simple example of a non-continuous element, but it is by no means the only possible non-continuous parameter. A non-continuous parameter is occasionally referred to as a set when there are other parameter display options employed (especially when a continuous parameter is used) because the non-continuous parameter separates the data into a series of discrete graphs viewed next to each other on the same screen. When a continuous parameter is used in conjunction with a non-continuous parameter each discrete graph contains all of the parameter values of the continuous parameter, making each of the separate graphs look like a set of parameter values.

Color Code

A color code is used for experimental parameters whose parameter values exist independently of one another, but are not unrelated to one another. When the genome browser is colored by parameter, GeneSpring will order the parameters values from top to bottom in the colorbar. Please refer

Copyright 1998-2001 Silicon Genetics

to “Color by Parameter” on page 3-33 for details. Parameter Values are listed in alphabetic or numerical order.

Each color represents a category (or set of categories). When coloring the browser display by parameter, each parameter-value defined as a condition is assigned a color and every data point described by that parameter is drawn in that parameter’s color. This can be referred to as *Color by Parameter*. Using this parameter display option means the browser display shows the same gene multiple times; the number of times a single gene is drawn is equal to the number of parameter values defined as conditions. When the browser display is colored using a color option other than Color by parameter, it is impossible to visually distinguish which parameter-value a particular gene line or gene point represents, although separate gene lines for each parameter-value defined as a condition are still drawn. Please refer to “Re-order the Parameters” on page 2-10 for details on how to change that order. Individual patients, or strain types, are variables commonly defined as color codes (conditions) because, although they are different parameter values, it is interesting to see them visually compared to one another. It is likely the expression patterns of individual patients with the same disease are going to react in a similar way under similar conditions, often it is when the expression patterns are not similar that the results are interesting. This is where graphs of parameter-values defined as color-coded conditions are useful as they allow you to easily compare varying conditions of the same gene.

Copyright 1998-2001 Silicon Genetics

Annotation Tools

The Annotations menu in GeneSpring allows you to update annotations, make gene lists based on annotations, and build gene ontology tables. You can annotate almost any data object in GeneSpring by adding notes in the various inspectors. Annotations can also be searched using the Find Gene feature in the Edit menu. See “Finding Genes” on page 3-4 for details.

Updating your Master Gene Table with GeneSpider

After you have loaded a new genome, you can make sure it contains the latest information from the genome databases on the World Wide Web by using GeneSpider. To use GeneSpider, you will need to have GenBank accession numbers in your master gene table. GenBank accession numbers are usually added to column 10 of the appropriate gene in the master gene table, separated by semicolons. For details on adding information to your master gene table see “Your Master Gene Table file” on page H-1.

To Update Annotations using GeneSpider

1. Select **`Annotations > GeneSpider`**. (Pre-4.1 users: Select **`Tools > GeneSpider`**). Choose one of four options:

   - **Update genes from Silicon Genetics:** Retrieves gene information from the Silicon Genetics Mirror Database. The mirror database caches information from GenBank, LocusLink, and UniGene to ease the load on the NCBI server and allow you to update faster. If a requested gene is not found in the mirror database, or if the information was cached more than 30 days ago, the mirror server will update the information from all three databases.
   - **Update genes from GenBank:** Allows you to retrieve information on genes from GenBank.
   - **Update genes from LocusLink:** Allows you to retrieve information from LocusLink.
   - **Update genes from UniGene:** Allows you to retrieve information from UniGene.

   The Update Genome window will appear.

2. Select the column containing GenBank accession numbers from the pull-down menu.
3. To update information in places where data already exists, select the **`Overwrite Existing Information`** checkbox. If you leave this box unchecked, GeneSpring will only add new information to blank fields. When you update annotations, GeneSpring creates a back-up file of the pre-update master gene table.
4. Choose where you wish to save your annotations. The default location is the master gene table you are currently using. For some genomes, you will have the option to save gene and non-gene information in different places. Updating from Silicon Genetics or GenBank will give you the option to retrieve sequence data. Updating from UniGene requires that you choose an organism from the pull-down menu, e.g. human, rat, mouse, zebrafish, cow, or frog.
5. Click **`Start`** to begin updating annotations.

Copyright 1998-2001 Silicon Genetics

Building a Simplified Ontology

New to GeneSpring 4.1 is the Build Simplified Ontology function, which builds a gene ontology list based on the Gene Ontology Consortium classifications. GeneSpring builds a hierarchical list from data found in all fields of the master gene table. The Build Simplified Ontology function places over 300 biologically meaningful groups in lists that can be compared and merged. By using these Gene Ontology lists you can study expression patterns of specific categories of genes by simply browsing through them.

Note: You cannot rename these gene lists, but you can update them.

To build a Simplified Gene Ontology list

1. Select **`Annotations > Build Simplified Ontology`**.
2. Name your folder.
3. Click **`OK`**. You will find your new Simplified Ontology list in the Gene Lists folder.

To make Gene Lists From Properties

To create lists based on annotations, see “Making Lists from Properties” on page 4-19.

Copyright 1998-2001 Silicon Genetics

Changing the Experiment Interpretation

The Change Experiment Interpretation window allows you to determine how an experiment is to be displayed. You can change the upper and lower bounds of the vertical axis of your graph, the mode used to represent your data, whether to turn on the global error model, how you would like to view each parameter, and which flagged measurements you wish to be displayed.

Changing an experiment interpretation is useful not only for customizing initial display settings, but also because statistical analysis techniques in GeneSpring are carried out based on how your data is characterized in the interpretation. Because of this, it can be valuable to set up more than one experiment interpretation, then perform analyses on each one to compare the results of statistical testing on data that has been grouped and characterized in different ways.

When you load your experiment GeneSpring automatically creates a Default Interpretation and an All Samples interpretation. The Default Interpretation is the first item listed under the experiment in the navigator. You will find it convenient to set up your most frequently used interpretation as your Default Interpretation. You can rename the Default Interpretation, but you cannot delete it. The All Samples interpretation makes all parameters non-continuous, so that each parameter is viewed and analyzed individually. The All Samples interpretation cannot be changed, renamed or deleted.

To change the Experiment Interpretation

1. Select **Experiments > Change Experiment Interpretation**. The Change Experiment Interpretation window will appear. (You can also right-click the genome browser in graph view and select **Options > Change Experiment Interpretation**.)
    - From the top pull-down menu, choose a data display mode for the vertical axis: **Ratio (signal/control)**, **Log of ratio** or **Fold Change**. The mode you choose will be used in such statistical procedures as Statistical Group Comparison, k-means Clustering, Self-organizing Maps, and Principal Components Analysis. See below for details on these modes. Choose the lower and upper bounds of the vertical axis in the fields provided.
    - If you do not wish to use the Global Error Model, deselect the **Use Global Error Model** checkbox. Using the Global Error Model allows you to produce a better estimate of precision. You can use these estimates in a number of analyses, including filtering and clustering. For information on the Global Error Model, see “Global Error Models Technical Details” on page N-1. For details on Color by Significance, see“Color by Significance” on page 3-33.
    - Depending on your instrumentation, you may have flags indicating the degree to which your data is reliable. If you have flags, choose from the **Use Measurements Flagged** pull-down menu to limit data based on these flags.
    - Choose a mode for each parameter: **Continuous Element**, **Non-continuous**, **Replicate** or **Color Code**. Note that if you choose Color Code, you must also select **Colorbar > Color by Parameter**. See below for details on these modes.
2. Name your interpretation and click **Save** to overwrite your current interpretation or **Save As** to create a new interpretation.

Copyright 1998-2001 Silicon Genetics

You will find saved interpretations by clicking on the relevant experiment in the Experiments folder of the navigator. You can delete an interpretation you have created by right-clicking over it in the navigator and selecting **`Delete`** from the pop-up menu.

Vertical Axis Modes

The default display is Ratio, where normalized intensity values are graphed on the vertical axis. In this mode, values range from zero to infinity.

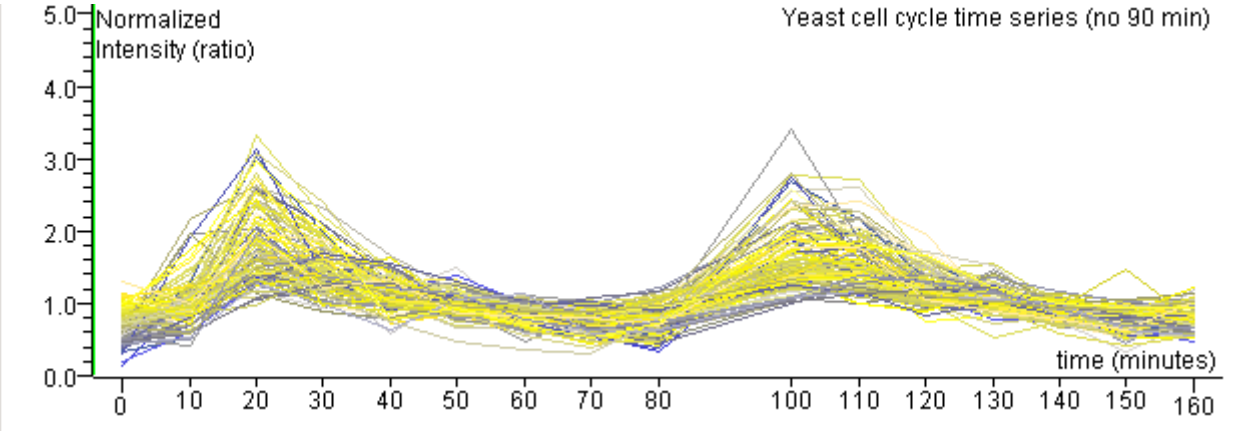


Figure 2-3 The gene list "like CLN1" graphed using the [signal/control] formula, The Y-axis is graphed from 0 to 5.

The ratio is determined by dividing the signal (raw data) by the control strength. (In a one-color experiment the control strength refer to the denominator used to normalize the raw data in a two-color experiment it is the control channel.) When data is reported as the signal divided by the control, it is assumed that all expression values are positive. The number 1 is considered normal expression; any expression value above one is overexpressed, and all underexpressed data is less than one, but greater than zero. This means that all underexpressed data appears flattened because it has to graphically fit between zero and 1, whereas overexpressed data takes up a much larger percentage of the graph (from 1 to positive infinity). Raw signal values that are negative (which is commonly the case in Affymetrix data) produce normalized values that are negative. (To deal with these negative values, see "The Affine Background Correction" on page 2-23.)

Log of Ratio

The Log of ratio mode graphs normalized values (i.e., the ratio of the signal to the control, not their logs), but spaces them logarithmically. The normal expression is 1. The Log of ratio interpretation solves the problem mentioned above under "Ratio", where all underexpressed data appears flattened because it has to graphically fit between zero and 1. In this mode underexpressed genes take up as much space visually as overexpressed genes. Logarithms of the expression ratios are used as the basis for statistical analysis.

Copyright 1998-2001 Silicon Genetics

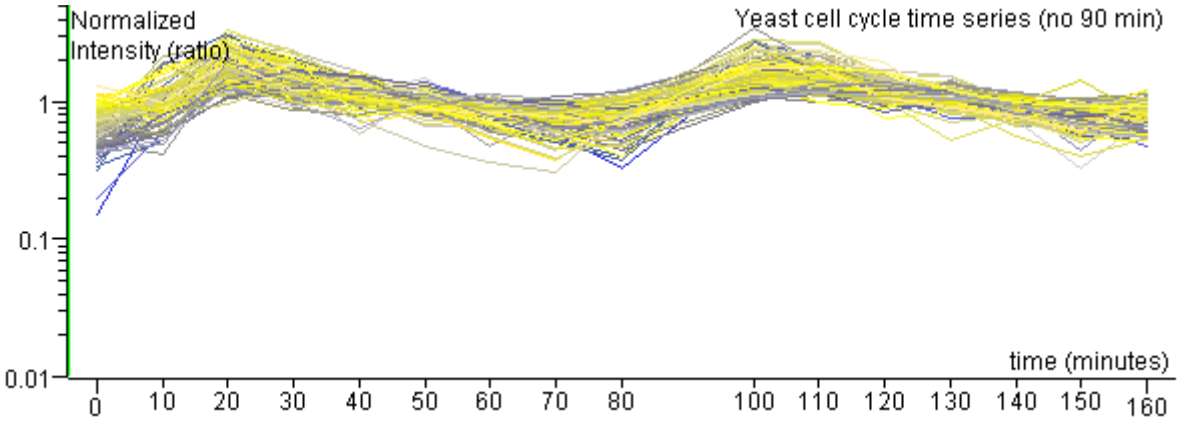


Figure 2-4 The gene list "like CLN1" graphed using the log ratio formula

Note that in Log interpretation, the lower limit of the vertical axis is 0.01. Any expression values below 0.01 are plotted as 0.01. Note also that when you export your data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01

Fold Change

Fold change mode creates a more balanced visual representation between over- and underexpressed genes than Ratio mode and emphasizes the increase and decrease of expression levels. For example, x1 would refer to normal expression, x2 to an expression level twice normal, and /2 to an expression level half normal. When using the upper or lower bound fields to change the vertical axis range enter either the ratio values in integers, or the fold change value (i.e., x4 or /4). Any integers you enter will be converted as in Table 2-1.

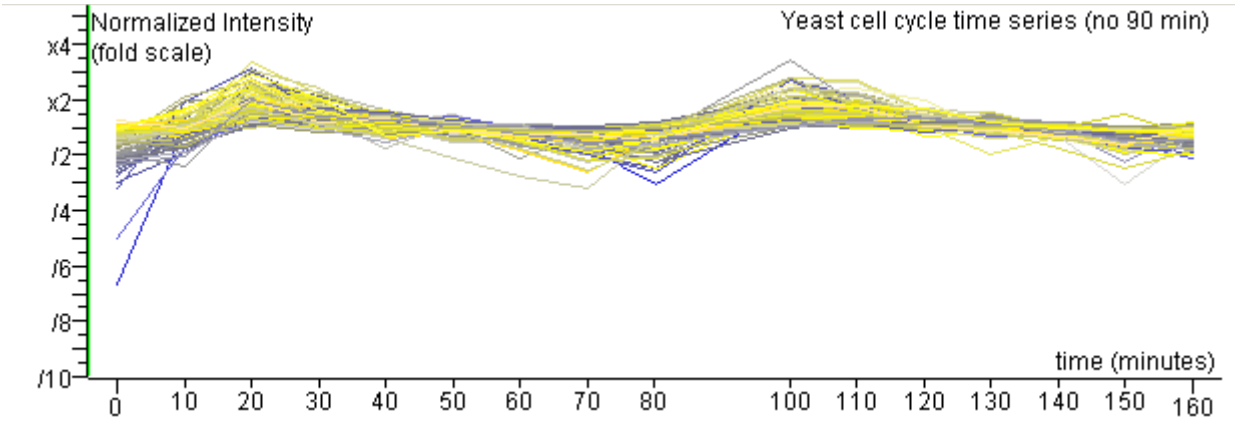


Figure 2-5 New Fold Change Image

Note that in Fold change interpretation, the lowest measured value is 0.01. Any values below 0.01 will be calculated as 0.01. The minimum display value is /10. Note also that when you export your data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01.

Copyright 1998-2001 Silicon Genetics

Ratio Numbers	Display
-5	/110
0	/110
.01	/100 (this is the lower cutoff
.25	/4
.33	/3
.5	/2
1.5	x1.5
3	x3
5	x3

Table 2-1 Fold Change

Parameter Display Modes

Continuous Element

Applicable only to Graph view, the Continuous Element mode shows parameter values existing on a continuum, where each point is connected with a line. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order. See “Parameter Display Options” on page 2-12 for details.

Non-Continuous

Applicable only to Graph view, Non-continuous mode shows parameter values existing independently of one another, where each value is represented as a discrete point. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order. See “Parameter Display Options” on page 2-12 for details.

Replicate

This mode applies to one of several experimental scenarios in GeneSpring:

- When you have one sample split across more than one chip.
- When you have multiple samples representing the same state.
- When samples from multiple tissues represent the same state.

Parameter defined as replicates are averaged together and appear as a single parameter.

Note that when the same gene occurs twice in the course of an experimental set, it is called a “repeat” and the measurements are averaged together. This cannot be changed.

Copyright 1998-2001 Silicon Genetics

Color Code

The Color Code mode colors genes by parameter. the number of times a single gene is drawn is equal to the number of parameter-values defined as conditions allowing you to easily compare varying conditions of the same gene. By default, parameter values are listed in alphabetic or numerical order. See "Parameter Display Options" on page 2-12 for details.

Experiment Normalizations

To normalize in the context of DNA microarrays means to standardize your data to be able to differentiate between real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales your data so that you can compare relative gene expression levels.

GeneSpring assumes that the data that you have entered is raw data that needs to be normalized. Note that if your data has been pre-normalized around a median other than 1, it may not be interpreted accurately during analysis. If your data is pre-normalized this way, please refer to "Use Constant Values" on page 2-24 or "Normalizing Each Sample to a Hard Number" on page G-7.

There are several ways to normalize your data in GeneSpring. Typically, you will want to do either one per-chip normalization together with one per-gene normalization or one per-spot normalization with one per-chip normalization. There are important exceptions to this, which are discussed below under the relevant normalization.

Note also that the order in which normalizations are performed is mathematically significant; GeneSpring performs them in the order in which they are listed here (and in the Experiment Normalizations window).

To get to the Experiment Normalizations window to assign normalizations, select `Experiments > Experiment Normalizations`.

Background Subtraction

To estimate background noise, some chips come with negative control spots that do not correspond to mRNA from the species under study. Even if your imaging software automatically subtracts background fluorescence, you may still want to tell GeneSpring to normalize to negative controls. The formula used here is:

(signal strength of gene A in sample X)
-(median signal of the negative controls in sample X)

To Subtract Background Noise

1. Create a negative control file by listing the names of your negative controls in the first column of a spreadsheet file and saving in tab-delimited text format.
2. Click the **`Use negative controls`** box.
3. Browse for the name of your negative control file.

Copyright 1998-2001 Silicon Genetics

Per-spot Normalization

If you are conducting a two-color experiment, you will probably want to do a per-spot normalization. The formula for this normalization is:

(signal strength of gene A in sample X)
(control channel value for gene A in sample X)

To Perform a Per-spot Normalization

1. Under *Per spot normalizations* choose either **`Use control channel to calculate ratio`** or **`Use control channel for trust`**, depending on whether or not your instrumentation has already calculated the ratio of the signals. The **`Use control channel for trust function`** tells GeneSpring to use the control channel to determine the saturation of the color of your genes.
2. In the **`Use values over`** box enter the value below which you do not trust the control signal (values below this cut-off will be thrown out).

Per-chip Normalizations

You will usually want to perform a per-chip normalization, which controls for chip-wide variations in intensity. This variation could be due to inconsistent washing, inconsistent sample preparation, or other microarray production or microfluidics imperfections. GeneSpring will not allow you to perform more than one per-chip normalization, as they all address the same issue.

If you have flags assigned to your data, select which data you would like used in your per-chip normalization from the **`Use genes marked`** pull-down menu.

Use Positive Control Genes

Some chips come with positive controls (mRNA from another genome or *housekeeping genes*, which are used to control for differences in the amount of exposure between samples. The formula for this difference is:

(signal strength of gene A in sample X)
(median signal of the positive controls in sample X)

To use Positive Control Genes

1. Create a separate positive control file by listing the names of your positive controls in the first column of a spreadsheet and saving in tab-delimited text format.
2. Under *Per chip normalizations* click **`Use positive control genes`**.
3. Browse to find your positive control file.
4. Enter a cutoff in the **`Use Values Over`** box telling GeneSpring not to do the normalization if the median of your chip is below this cutoff.

Copyright 1998-2001 Silicon Genetics

- **One caveat regarding normalizing to positive controls:** This normalization will not control for variations in the total harvest of mRNA across samples. If you are concerned about this variation, you may want to instead normalize to the distribution of all genes.

Normalizing to the Distribution of All Genes

The most common way to control for systematic variation is by normalizing to the distribution of all genes. The formula for this is:

(signal strength of gene A in sample X)
(specified percentile of all of the measurements taken in sample X)

To Use Distribution of All Genes

1. Under *Per chip normalizations* in the Experiment Normalizations window click **`Use distribution of all genes`**.
2. Typically you will use the default percentile (50th).
3. Enter a cutoff in the **`Use Values Over`** box telling GeneSpring not to do the normalization if the median of your chip is below this cutoff.

- **One caveat:** This sort of normalization assumes that the median signal of the genes on the chip stays relatively constant throughout the experiment. If the total number of expressed genes in the experiment changes dramatically due to true biological activity (causing the median of one chip to be much higher than another), then you have masked your true expression values by normalizing to the median of each chip. For such an experiment, you may want to consider normalizing to something other than the median or you may want to instead normalize to positive controls.

Region Normalization

If you have more than one chip assigned to a sample, and you would like to normalize them separately, you can do a region normalization. You can also do a region normalization if you would like to normalize a region of a particular chip separately from the rest of the chip. To do this, you will need to load your data through the Experiment Wizard (see “Region Normalization” on page G-15). If after loading your data you would like to change the way your regions are designated, you can do so in the Experiment Normalizations window under Region Designators.

The Affine Background Correction

If negative values form a large fraction of your data set, GeneSpring may automatically do what is known as the affine background correction. If a large percentage of your data is negative, normalization can be a problem; for instance, the median, which GeneSpring divides your data by in Use Distribution of All Genes, can be very small or even negative.

In such cases, GeneSpring will readjust the background level for your data by adding a constant to all raw control strengths such that the 10th percentile is set equal to 0. The affine background correction is applied only when the 10th percentile is more negative than the median of the data is positive. You will get a warning message when loading your data if the correction is applied.

Copyright 1998-2001 Silicon Genetics

Also, in the Gene Inspector, control strengths adjusted by this correction are flagged with asterisks.

To tell GeneSpring If and When to Apply the Affine Background Correction

The **Options** pull-down menu in the Experiment Normalization window allows you to do this.

- **Use simple ratio:** Tells GeneSpring to never use the affine background correction. If the control value is negative GeneSpring will produce a warning message and will not do the normalization.
- **Use ratio with background correction:** Tells GeneSpring to always use the affine correction. You will only want to select this option if no background subtraction has been performed on your data, as it forces the 10th percentile to be 0 (as if it were considering 10 percent of the data background). As nearly all image analysis software has already done background subtraction, this should be a rarely used option.
- **Use background correction if needed:** Tells GeneSpring to use the affine correction as needed to compensate for negative values.

Use Constant Values

If you are using a technology that calculates its own number for normalization you will want to use constant values. For instance, Affymetrix's Global Scaling™ centers your data around 2500; in this case you would need to normalize your data to 2500 to center it around 1.

(signal strength of gene A in sample X)
(hard number in sample X)

To use Constant Values

1. Under *Per chip normalizations* click **Use constant values**.
2. Specify the hard number for each of your samples.

Copyright 1998-2001 Silicon Genetics

Per-gene Normalizations

Normalize to Median For Each Gene

This per-gene normalization accounts for the difference in detection efficiency between spots. It also allows you to compare the relative change in gene expression levels, as well as display these levels in a similar scale on the same graph. GeneSpring uses the following formula to normalize to the median for each gene:

(signal strength of gene A in sample X)
(median of every measurement taken for gene A throughout your experiment)

To Normalize to the Median For Each Gene

1. Under *Per gene normalizations* click **`Use median for each gene`**.
2. Enter a number that is an estimate of the lowest signal value that you trust. If a median value falls below this cut-off, the program will instead divide by the cut-off.

GeneSpring will not allow you to do this normalization and normalize to sample(s), as they address the same issue.

Normalizing to Sample(s)

In normalize to sample(s) each gene is divided by the intensity of that gene in a specific control sample or by the average intensity in several control samples. The formula for this is:

(signal strength of gene A in sample X)
(signal strength of gene A in the control sample[s])

Or,

(signal strength of gene A in sample X)
(average signal strength of gene A in several control samples)

To Normalize To Sample(s)

1. Under *Per gene normalizations* click **`Use sample(s)`**.
2. Indicate the numbers of the control samples (sample numbers are listed under **`Experiments > Change Experiment Parameters`**). Multiple experiment numbers must be separated by commas (e.g., 1,2). Ranges of experiment numbers can be indicated by a dash (e.g., 1-3,5). You also have the option of normalizing subsets of your samples to the mean of specific subsets of control samples. For more information, click the Use sample(s) **`Help`** button.
3. Specify a cutoff for the denominator in the above formula. The cutoff is used on measurement values that have been partially normalized in previous normalization steps, so this should be a small number, like 0.01. If the denominator falls below the cutoff, and the numerator is above the cutoff, the denominator used for the above formula will be .01. If both the numerator and

Copyright 1998-2001 Silicon Genetics

the denominator fall below the cutoff, this measurement will not be included in the normalization.

GeneSpring will not allow you to do this normalization and normalize to the median of each gene, as they address the same issue.

Miscellaneous

Regarding **normalizing merged/split experiments**, you have the option of starting with the original normalizations or discarding these and starting with raw data. The default setting starts you with the original normalizations. To start with the raw data, deselect the **`Start with normalized data`** checkbox in the Experiment Normalizations window.

You can assign a minimum value for your measurements. Any measurements that fall below this minimum value will be assigned the minimum value. To assign a minimum value:

1. Check the **`Make Minimum Value`** box under *Miscellaneous*.
2. Enter a minimum value in the field to the right.

Global Error Models

Using the Global Error Model

The error model has changed significantly in GeneSpring 4.1, and now separate estimates of two different kinds of random variation are used to estimate the variability in gene expression measurements:

- **Measurement variation:** This comprises the lowest level of variation, corresponding to the variation of the measurement of a gene on a single chip around the true value that would be achieved by a perfect measurement of the expression level of the gene for that sample.
- **Sample-to-sample variation:** This is the variation between samples in the same condition. This represents biological or sampling variability, such as variability between multiple subjects in a condition, between multiple physical samples for an experimental subject or patient, or between multiple hybridizations of a physical sample. GeneSpring can represent any one of these kinds of variability, depending on the types of replicate samples you have specified in your interpretation and in the error model dialog. GeneSpring assumes all replicate samples in the same condition correspond to one kind of variability.

The ability to estimate measurement and sample-to-sample variation in microarray-based experiments is often compromised by the fact that the cost (in both time and materials) of performing large numbers of replicate experiments is quite high. If the global error model is turned on, GeneSpring accounts for error instead by assuming that the amount of variability is a function of the control strength within all the measurements for a single experimental condition. The advantage of making this assumption is that the number of measurements used to estimate the global error is equal to the total number of genes on any given chip.

Copyright 1998-2001 Silicon Genetics

In addition, measurement precision information supplied by the scanner software or independently by the user can be loaded into GeneSpring via the "Signal Precision" column type in the column editor. The value given in this column is interpreted as the standard deviation of the raw measured value.

The sample-to-sample variability includes the effect of both types of variation, and the statistical separation of these effects is called *variance components analysis*. The GeneSpring Global Error Model performs this variance components analysis, and uses the estimates of these two components of variation to accurately estimate standard errors and compare mean expression levels between experimental conditions.

When you turn the Global Error Model on the Error Model is used as the basis for:

- standard deviation, representing the variability of individual population members
- standard error, representing the precision of the mean of the gene expression measurements in the condition with respect to the true condition mean
- error bars corresponding to standard deviation or standard error in the Graph view and Gene Inspector
- t-test p-value, representing the statistical test of differential expression for a specific condition
- color by significance, coloring according to the t-value from the t-test of differential expression
- tests between condition means using the Statistical Group Comparisons filter, if the error model option is chosen.

To turn on the Global Error Model

1. Select **Experiments > Error Models**. The Error Models window will appear.
2. If you have replicates for each condition, check the Replicates box and select parameters to treat as replicates. Click **OK**. If not, check the Deviation from 1.0 box. Click **OK**.
3. Select **Experiments > Change Experiment Interpretation**. The Change Interpretation window will appear.
4. Click the box marked **Use Global Error Model**.
5. Click **Save** to save as part of your current interpretation or **Save As** to create a new interpretation.

Copyright 1998-2001 Silicon Genetics

Technical Details

The two-component model for estimating variation from control strength is known as the Rocke-Lorenzato model. The two components are an absolute error component that dominates at low measurement levels, and a relative error component that dominates at high measurement levels. The formula for the error model for raw (pre-normalization) expression levels can be written as:

$$\sigma_{RAW} = \sqrt{a^2 + b^2 S^2}$$

where σ_{raw} is the measurement standard error of the raw expression data, *S* is the measurement level (control strength), and *a* and *b* are the fitted coefficients of the model.

Expressed in terms of the normalized expression levels, which are the result of dividing raw expression levels by control strength, the standard errors can be written as:

$$\sigma_{RAW} = \sqrt{a^2 + b^2 S^2}$$

Before fitting the error model, the genes are ordered by their control strengths. A median variance and median control strength is calculated for each non-overlapping set of eleven genes. If replicates are used, this variance is the standard error of the samples in the current condition. If the "deviation from 1" option is selected, error is approximated by using the median deviation from 1.0. The goal in this step is to remove outliers (when replicates are being used) and to disregard genes whose high or low expression level is the result of biological activity. In the absence of replicates the working assumption is that the vast majority of the genes do not change over the conditions in the experiment, and thus deviation from one represents error in a gene whose expression level changes little over the course of the experiment. Then an iteratively reweighted linear regression of variation or squared deviation versus squared control strength is fitted to estimate the parameters.

Estimation of the 2-level variance components model is done by the method of moments. In order to eliminate negative estimates of variance components, within-sample variation is taken as a lower bound on total between-sample variation. Different sources of information in the analysis are weighted by their appropriate statistical degrees of freedom. Precision estimates based on replicate genes or samples are assigned degrees of freedom equal to the number of replicates minus one. User-supplied precision values, if available, are assigned 1 degree of freedom. Cross-gene error models, if used, are assigned an equal number of degrees of freedom as the direct variability estimates for that gene. Between-sample analyses are done according to the interpretation mode (ratio, log, fold). Within-sample variability is calculated in terms of normalized ratio expression, and translated as necessary to the interpretation mode by use of the delta method.

Copyright 1998-2001 Silicon Genetics

Results of the variance components analysis are used to estimate standard deviations and standard errors, according to the grouping of samples into conditions as specified by the experiment interpretation. Two different types of interpretation affect the assumed context of the calculation:

- **Single-sample interpretation:** If all conditions contain only one sample (for instance the. "All Samples" interpretation), precision calculations are based solely on the estimated within-sample measurement variation. The error bars, standard deviations, and standard errors represent the variability of all possible measurements on this specific sample.
- **Multi-sample interpretation:** If at least one condition contains multiple samples, precision calculations for all samples are based on the combined within-sample and between-sample variation, and error bars, standard deviations, and standard errors represent the variation of measurements of samples representing the population of all possible samples in the condition.

In a multi-sample interpretation, if no replicate samples are available for a specific condition, then no error calculations are made and no error bars are shown, since there is no information available on the variability of that condition.

References

Rocke, D.M., and S. Lorenzato. 1995. A two-component model for measurement error in analytical chemistry. Technometrics 37:176-184.

Milliken, G. A. and Johnson D, E. (1984) *Analysis of Messy Data*, Volume 1: Designed Experiments Wadsworth, Inc. Belmont, California.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, John Wiley and Sons, New York.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin 2: 110-14.

Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Chapter 3 Viewing Data in GeneSpring

Using Genome Browser

The large panel in the center of the GeneSpring window is the genome browser, which graphically displays information about the genes in the selected gene list. The genome browser often presents so much information that individual genes and gene names are not visible. To look more closely at fewer genes you can zoom in and pan around.

Zooming In

You can enlarge a region of the screen by "zooming in".

1. Click and drag a rectangle across the region you wish to enlarge.
2. Release the cursor. Repeat steps 1 and 2 until you reach the desired magnification level.
3. To undo a zoom, type **`Ctrl+Z`**.

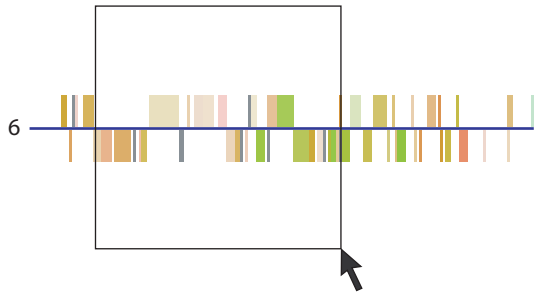


Figure 3-1 Zooming

To return directly to the unmagnified state, do one of the following:

- Select the **`View > Zoom Fully Out`** option.
- Type **`Ctrl + Home`**.

Panning

If you have zoomed in and need to view genes that are not visible in the genome browser but are nearby, you can pan in any direction.

To pan, do one of the following:

- Use the arrow keys to move in the desired direction.
- Use the Page Up or Page Down keys to travel one screen's distance up or down.

Copyright 1998-2001 Silicon Genetics

Changing Genome Browser Elements

To change genome browser elements, right-click on the genome browser to select any of the following items in the **Options** submenu:

- **Change Vertical Axis Range—**Allows you to change the range of the vertical axis.
- **Show/Hide Timeline—**Allows you to show or hide the timeline.
- **Show/Hide Horizontal Label—**Allows you to show or hide the label on the horizontal axis.
- **Show/Hide Vertical Label—**Allows you to show or hide the label on the vertical axis.
- **Label Vertical Axis at Top/Label Vertical Axis on Side—**Gives you the option of placing the vertical axis label on the side of the vertical axis or on top.
- **Show/Hide Experiment Name—**Allows you to show or hide the experiment name in the top right corner.

Error Bars

You have the option of using error bars in the Graph and Scatter Plot views. To turn the error bars on, right-click in the genome browser and select **Error Bars > Show Error Bars**. The error bars will be visible in the Gene Inspector as well as in the main GeneSpring window.

You can choose one of the following three kinds of error bars:

- Standard Error
- Standard Deviation
- Minimum/Maximum Value of Each Gene

To access one of these options, right-click on the genome browser and select the **Error Bars** submenu. Note that to select an error bar type you must first have selected **Error Bars > Show Error Bars**.

Copyright 1998-2001 Silicon Genetics

Splitting Windows

The Split Windows feature allows you to view several classifications or lists of genes separately in the genome browser. If you switch to another view in the View menu, the window will remain split. While viewing split screens you can zoom, pan and make changes in the experiment interpretation the same way you do with unsplit screens.

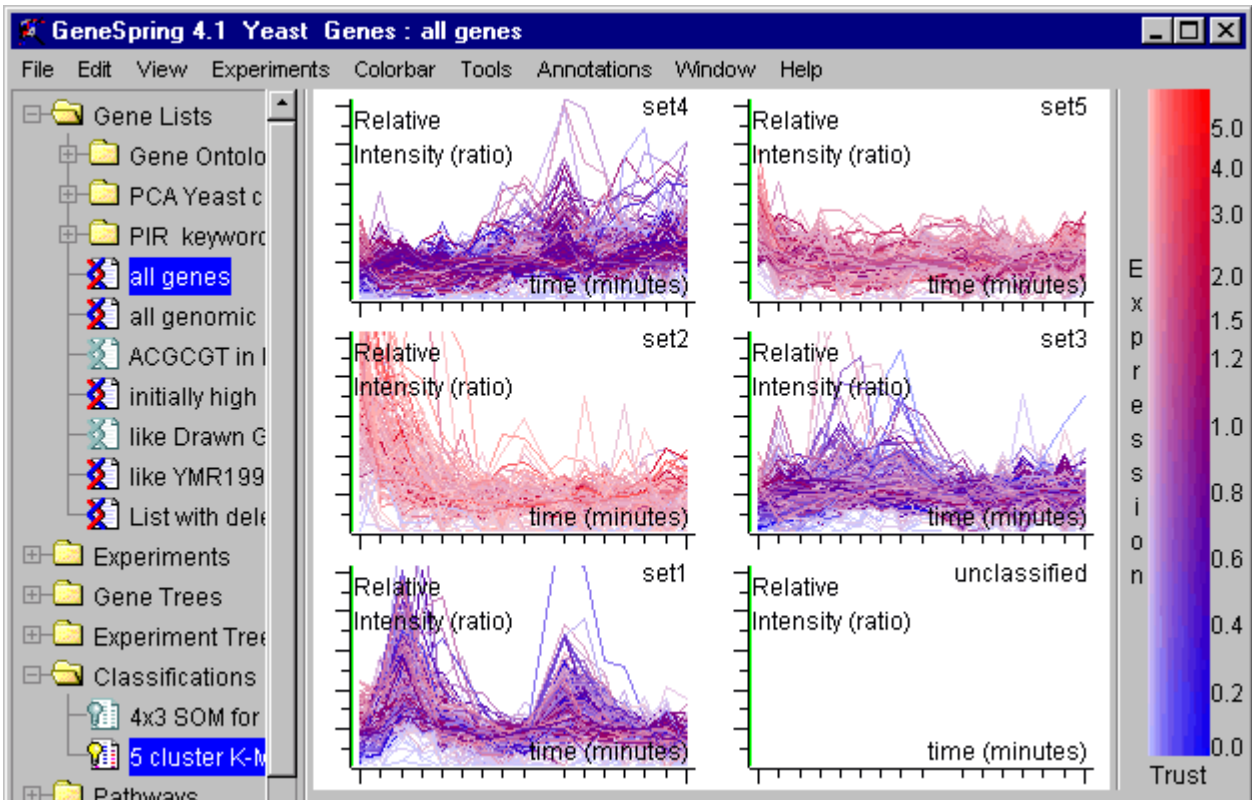


Figure 3-1 Example of a k-means clustering

In Figure 3-1, the example represents a k-means clustering, colored by expression values. Note the list name and number of genes shown in the upper right corner of each small screen. In this instance, the names are set numbers from the original k-means clustering.

To Split a Window

1. Right-click a gene list folder or classification in the navigator and select **`Split Window`**. A submenu will appear.
2. Select from one of the following display options:
   - **`Horizontally`** – to divide the window into columns
   - **`Vertically`** – to divide the window into rows
   - **`Both`** – to create a grid

To unsplit a window, select **`Split Window > Neither`** or **`View > Unsplit Window`**.

Copyright 1998-2001 Silicon Genetics

Displaying a Gene List

To display a gene list:

1. Right-click on the gene list you wish to view in the Gene List folder in the navigator. A submenu will appear.
2. Select **Display List**.

Displaying a Gene List as a Secondary List

1. Display a gene list as outlined above, then right-click above the gene list you wish to view as your secondary gene list. A submenu will appear.
2. Select **Display As a Second List**.

To remove the secondary gene list, go to the **View** menu and select **Remove Secondary Gene List**.

Finding and Selecting Genes

The Find Gene function allows you to quickly find a gene when using a view where individual genes are not easily distinguished.

Finding Genes

1. Go to **Edit > Find Gene**. The Find Gene window will appear.
2. Type a keyword, systematic name or common name of a particular gene in the Find Gene window text box.
3. Click **OK** or press the **Enter** key.

If GeneSpring does not recognize the word you typed in you may get an error message.

In some views, the genome browser will zoom in on the “found” gene. This gene will be automatically selected.

If your search results in more than one matching gene, GeneSpring will provide you with a list to choose from. To reduce the number of matches, type a whole word into the Find Gene box. A partial word like “prot” will result in a list with every instance of the string “prot” in it. The more specific you can make your search string the fewer numbers of genes you will have to sort through in the Multiple Results window.

Copyright 1998-2001 Silicon Genetics

Selecting Genes

Often you will need to select a gene or group of genes in order to identify gene names or quickly access genes you are working with.

To Select a Single Gene

- Click once on any line or square representing a gene. The name of this selected gene will appear in the upper right corner of the genome browser.
- Double-click a gene to bring up the Gene Inspector window (see "Gene Inspector" on page 3-37) or use **Ctrl+I** for a selected gene. This works on genes represented graphically in the genome browser and on gene names found in lists.

Tip: To select a gene in the genome browser, first zoom in on it.

To Select Multiple Genes

- Click once on any line or square representing a gene. Hold down Shift to add more genes. (Clicking a selected gene while holding **Shift** deselects that particular gene.)

  Or,

- **Shift** and drag your mouse across genes you would like to select. You will see a box appear as you drag. When you release the mouse, the selected genes will be highlighted.

When several genes are selected, no gene names appear in the genome browser.

If some selected genes do not appear in the current view, the upper right corner of the genome browser will display the message "Some selected genes not shown".

Click anywhere in the browser to deselect genes.

List Inspector

Right-clicking over a list icon in the navigator will bring up several options including **Inspector**. Selecting the **Inspector** command will open a List Inspector window displaying the common and systematic names of all the genes in the gene list currently being displayed in the genome browser. You can select one of the listed genes (by double-clicking) for closer inspection. For more information on this window, see "List Inspector" on page 3-44.

Copyright 1998-2001 Silicon Genetics

Showing/Hiding Window Display Elements

You have the option of showing or hiding many of the elements in the GeneSpring window. To change the visibility of these elements, select **`View > Visible`** and choose one of the following options:

- **Picture—**Shows or hides the optional picture at the bottom right corner of the window
- **Animation Controls—**Shows or hides the slider and the Animate check box at the bottom of the window (hiding this check box does not disable the Animation feature)
- **Magnification—**Shows or hides the Magnification feature and the **`Zoom Out`** button at the bottom of the window (hiding the **`Zoom Out`** button does not disable the **`Zoom Out`** menu option)
- **Secondary Picture—**Shows or hides your secondary picture when you are viewing two gene lists or experiments simultaneously in the genome browser
- **Secondary Animation Controls—**Shows or hides the secondary Animation Controls check box and slider when you are viewing two gene lists or experiments simultaneously
- **Navigator—**Shows or hides the navigator panel
- **Hide All—**Hides everything in the window except the genome browser
- **Show All—**Shows all elements
- **Hide All in All Windows—**Hides everything in all windows except the genome browser
- **Show All in All Windows—**Shows all elements in all windows

Copyright 1998-2001 Silicon Genetics

Graph View

The Graph view allows you to visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a line. To get to the Graph view, select **`View > Graph`**.

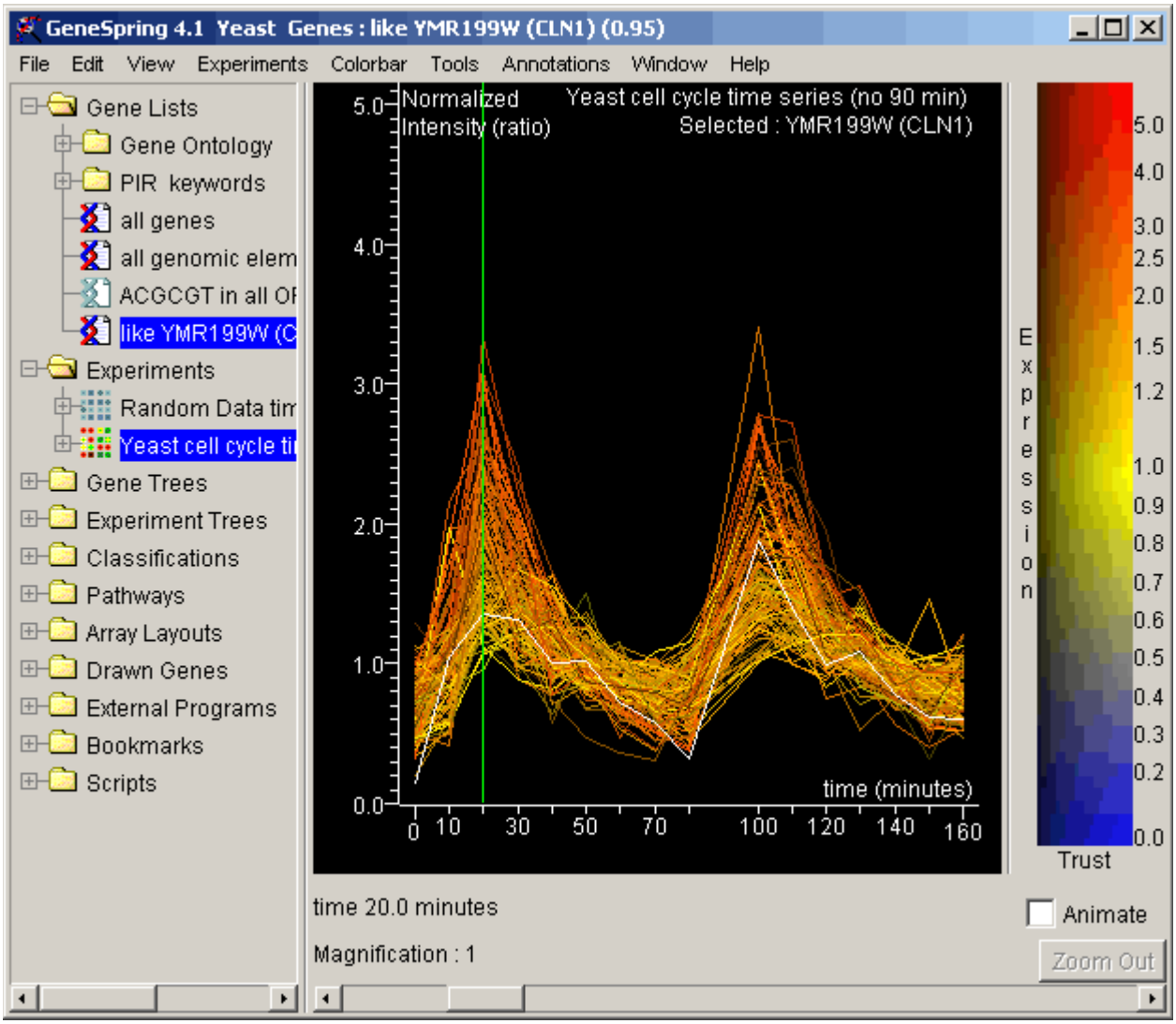


Figure 3-2 The Graph View

Figure 3-2 shows the genes in the "like YMR199W(CLN1)(0.95)" list in Graph view. The gene in white has been selected; its name appears in the upper right-hand corner of the genome browser, underneath the title of the experiment.

Copyright 1998-2001 Silicon Genetics

Bar Graph View

The Bar Graph view allows you to visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a vertical bar. To switch to Bar Graph view, select **`View > Bar Graph`**.

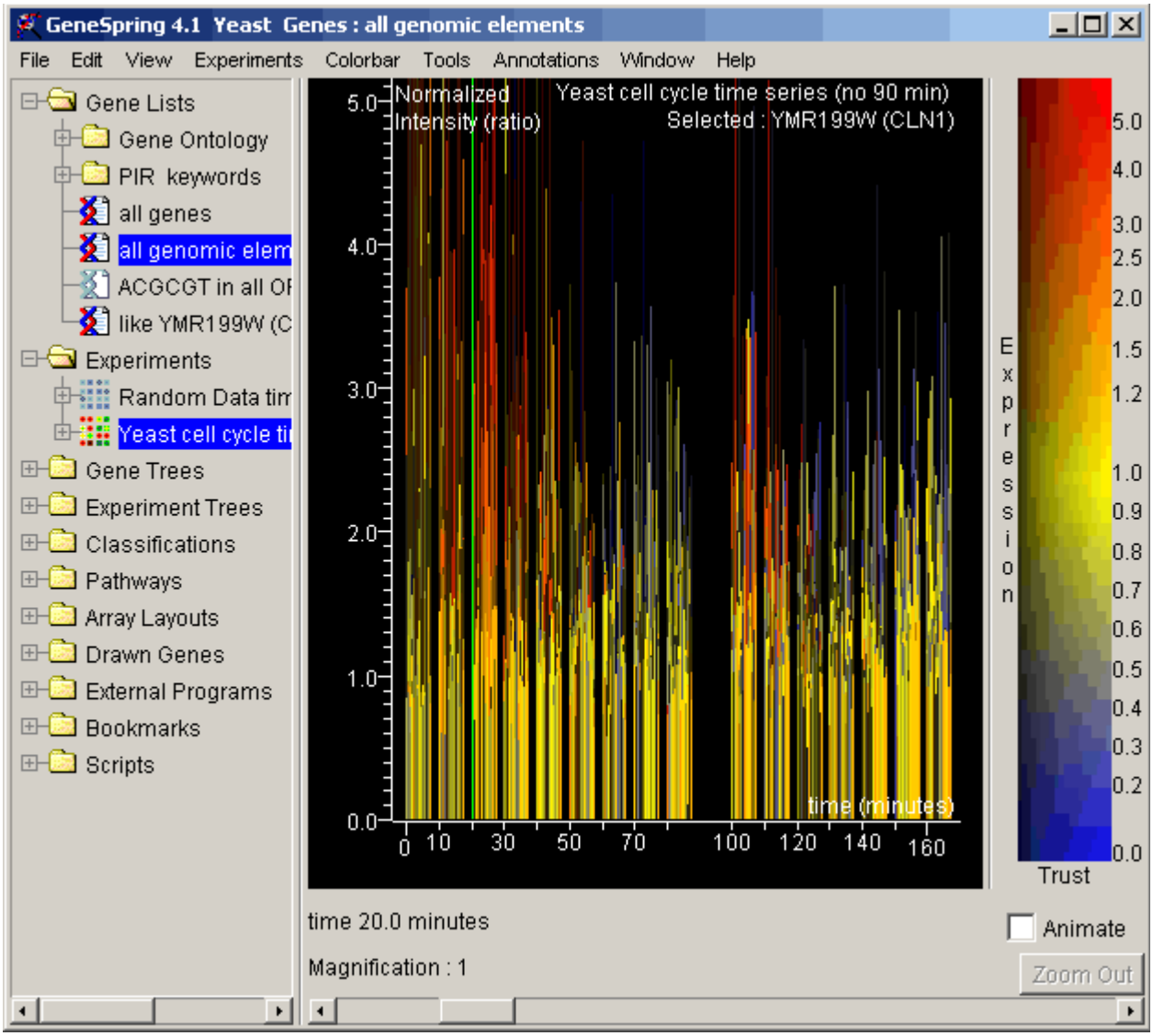


Figure 3-3 The Bar Graph view

Figure 3-3 shows a Yeast cell cycle time series in Bar Graph view.

Copyright 1998-2001 Silicon Genetics

Classifications View

This view allows you to visualize an experiment or a set of experiments by organizing the genes according to previously defined categories.

To use Classification view

1. Select a gene list.
2. Classify the genes using one of two methods:

   a. Right-click a subfolder in the Gene Lists folder and choose **`Use as Classification`** from the resulting pop-up window.

   b. Select a previously created classification from the Classifications folder in the navigator (see "Clustering and Characterizing Data in GeneSpring" on page 5-1).

Color genes by your chosen classification:

1. Select **`Colorbar > Color by classification`**.
2. Right-click a subfolder and select **`Use as Coloring`**.
3. Right-click an existing classification in the Classifications folder and choose **`Set as coloring scheme`**.
   For more information on coloring see "Changing the Coloring Scheme" on page 3-31.

You can also see how many genes have no data by noting how many genes are greyed out. If you switch to other views you can return via **`View > Classification`** (automatically selected by classifying a list using the methods above).

Note: If you select Classification from the View menu without specifying a classification method, the genome browser will display the genes without any classification.

Copyright 1998-2001 Silicon Genetics

Physical Position View

This Physical Position display allows you to see an experiment or a set of experiments by organizing the genes according to their physical position (when the gene loci are known and loaded into GeneSpring) within the DNA sequence the organism. Select **`View > Physical Position`**. The Physical Position view works for any organism whose mapping data is at least partially available. An illustration of what Physical Position View looks like for humans is given Figure 3-5. For organisms already sequenced, the physical position views will look more like yeast (illustrated in Figure 3-4).

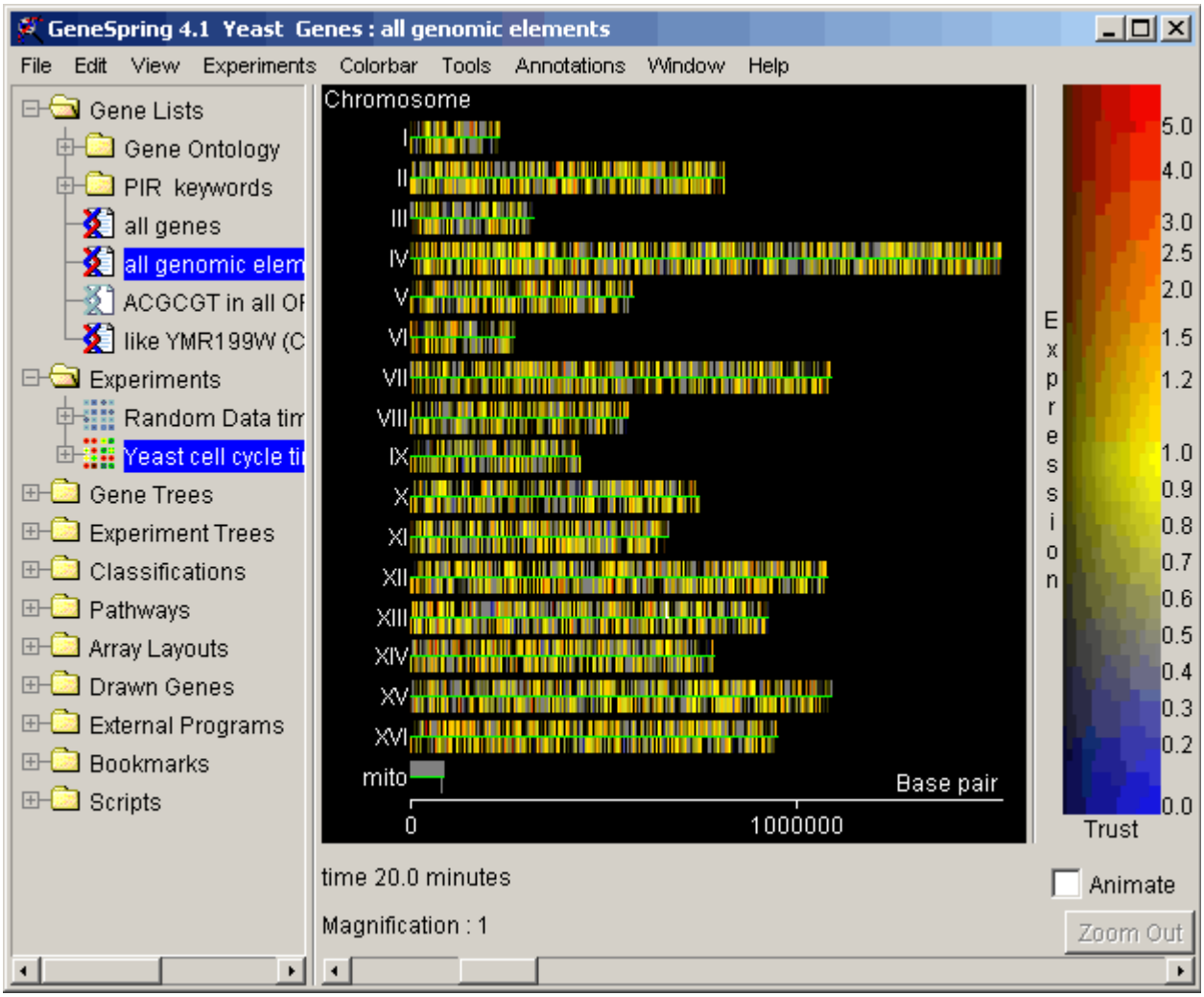


Figure 3-4 The Physical Position view

The Physical Position view for yeast is also discussed in *GeneSpring Basics Instructional Manual* 1.3 “Physical Position Display” on page 1-5.

At greater magnification, you can see the base pairs.

Copyright 1998-2001 Silicon Genetics

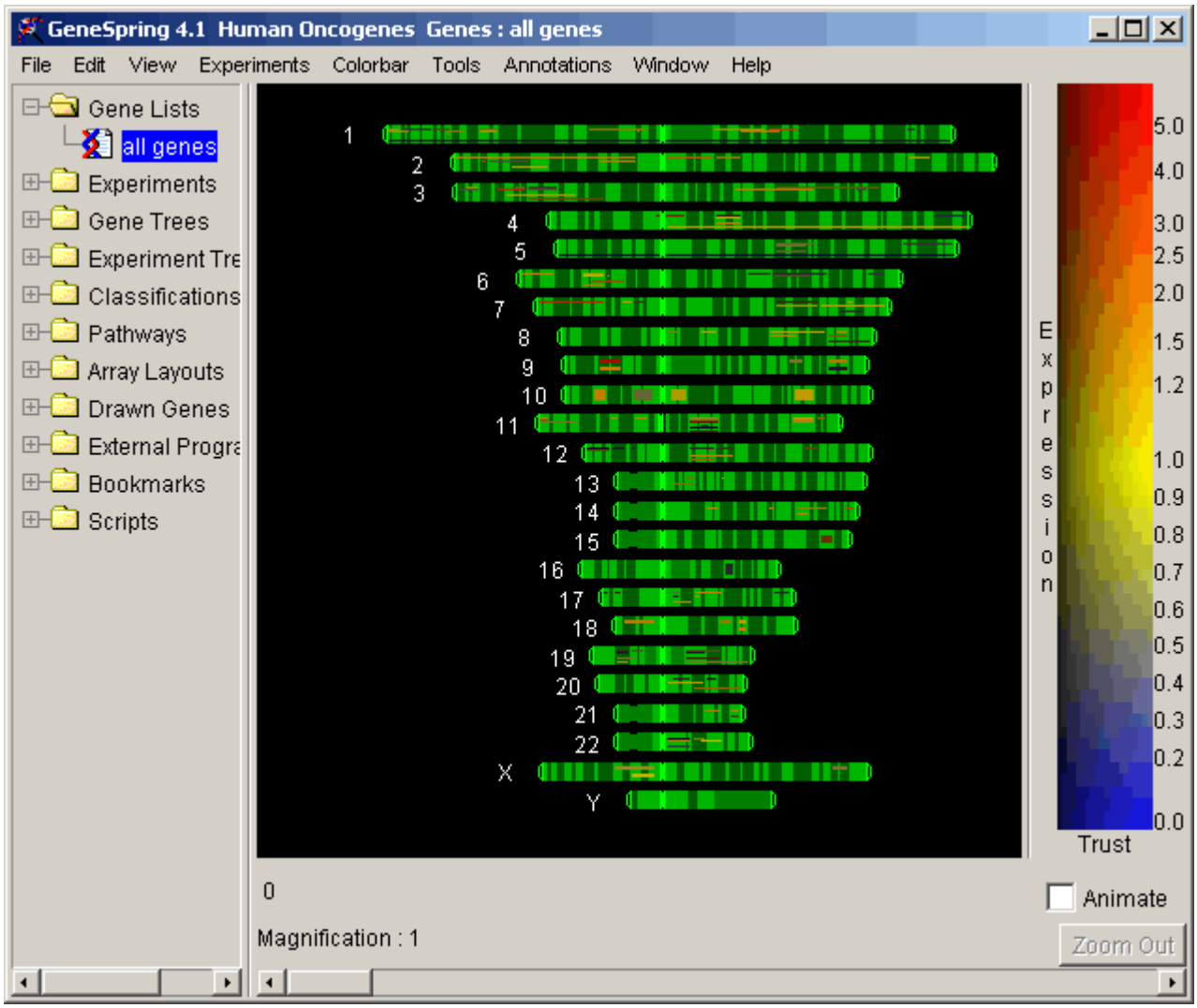


Figure 3-5 Physical position view for human oncogenes

Copyright 1998-2001 Silicon Genetics

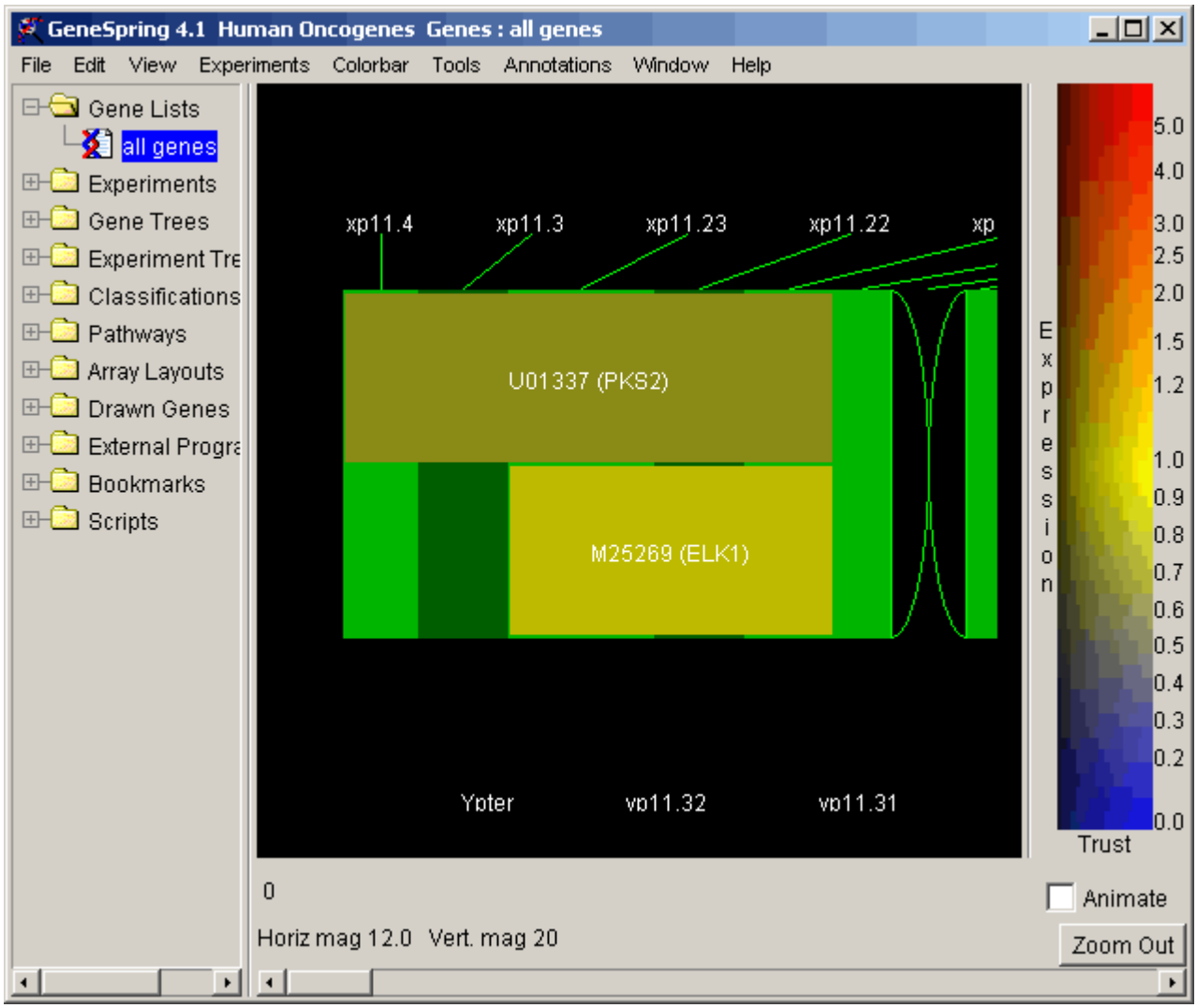


Figure 3-6 Zooming in for a closer look at the Y chromosome

At high magnifications the labels associated with the chromosome's cytogenetic bands are displayed.

Copyright 1998-2001 Silicon Genetics

To use the Load Sequence command

In GeneSpring versions 4.0 and later the default setting of the program is to load the sequence information if available. If you have an old version of GeneSpring and cannot update it (please refer to "Update GeneSpring" on page A-2), please follow these directions.
The Load Sequence command is only applicable for sequenced organisms. Loading the nucleic acid sequence allows you to magnify a section of the physical position view to the point where the nucleic acid sequence is displayed. Loading the sequence also allows you to take advantage of GeneSpring's other sequence-based features such as **Tools > Find Potential Regulatory Sequences**. Loading the nucleic acid sequence can be done in a number of ways.

Method 1 takes immediate effect.

1. Right-click while the cursor is in the black genome browser. A menu will appear.
2. Select **Options > Load Sequence**.

A window saying *Please wait while nucleic acid sequence is loaded* will appear. After the loading is complete it is possible to zoom in and see the nucleic acid sequence of a particular gene.

The sequence will be shown in the magnified genes. However, this information is not saved, so when you exit GeneSpring and re-open you will need to reload the nucleic acid sequence.

If you would like the sequences to always be readily available, you must change the defaults through the Preferences window. You may choose to make the load sequence feature automatically load with the program. Again, please note that this applies to version 4.0 and earlier.

Method 2 takes effect in your next GeneSpring session:

1. Select **Edit > Preferences**. The GeneSpring Preferences window will appear.
2. Select **Data Files** from the drop-down at the top of the window.
3. Select the **Load Sequence** checkbox.
4. Click the **OK** button at the bottom of the window.
5. Close and restart GeneSpring. (Or, you can select **File > New Window**.)

Changing the defaults in the Preferences window will not initiate the load sequence feature in your current session, but it will change future initial loading practices. The nucleic acid sequence can also be loaded as a side effect of using **Tools > Find Regulatory Sequences**. For more information on this particular feature, see "Regulatory Sequences" on page 4-26.

To Show ORF direction/Ignore ORF direction

A gene is represented visually by a colored line or upon higher magnification a colored rectangle. The rectangle's position relative to the chromosome line determines the direction of the ORF. A gene below the chromosome line has a reading direction opposite to the direction chosen by the sequencers, and the sequence is read backwards. You can choose to display this distinction between which direction a gene is read (**Show ORF direction**) or to have no distinction between genes (**Ignore ORF direction**). To invoke either of these options:

Copyright 1998-2001 Silicon Genetics

1. Right-click while the cursor is in the genome browser. A menu will appear.
2. Go to the **Options** submenu.
3. Select the **Ignore ORF direction** command or the **Show ORF direction** command.

To Show complementary bases/Just show one strand of bases

Show complementary bases allows both of the complementary nucleotides to be shown while viewing the nucleic acid sequence in the physical position view. Conversely, *Just show one strand of bases* shuts this feature off and only views the Watson strand of the sequence. To invoke either of these options:

1. Right-click while the cursor is in the genome browser. A menu will appear.
2. Select **Options > Just show one strand of bases** or **Show complementary bases**.

Copyright 1998-2001 Silicon Genetics

Scatter Plot View

The Scatter Plot view is useful for examining the expression levels of genes in two distinct conditions, samples, or normalization schemes. For instance, you can use the scatter plot to identify genes that are differentially expressed in one sample versus another. A scatter plot can also be used to compare two values associated with genes in two gene lists. Such associated values might include the relative contribution of principal components as determined from principal components analysis, or two similarity scores from the Find Similar function in the Gene Inspector.

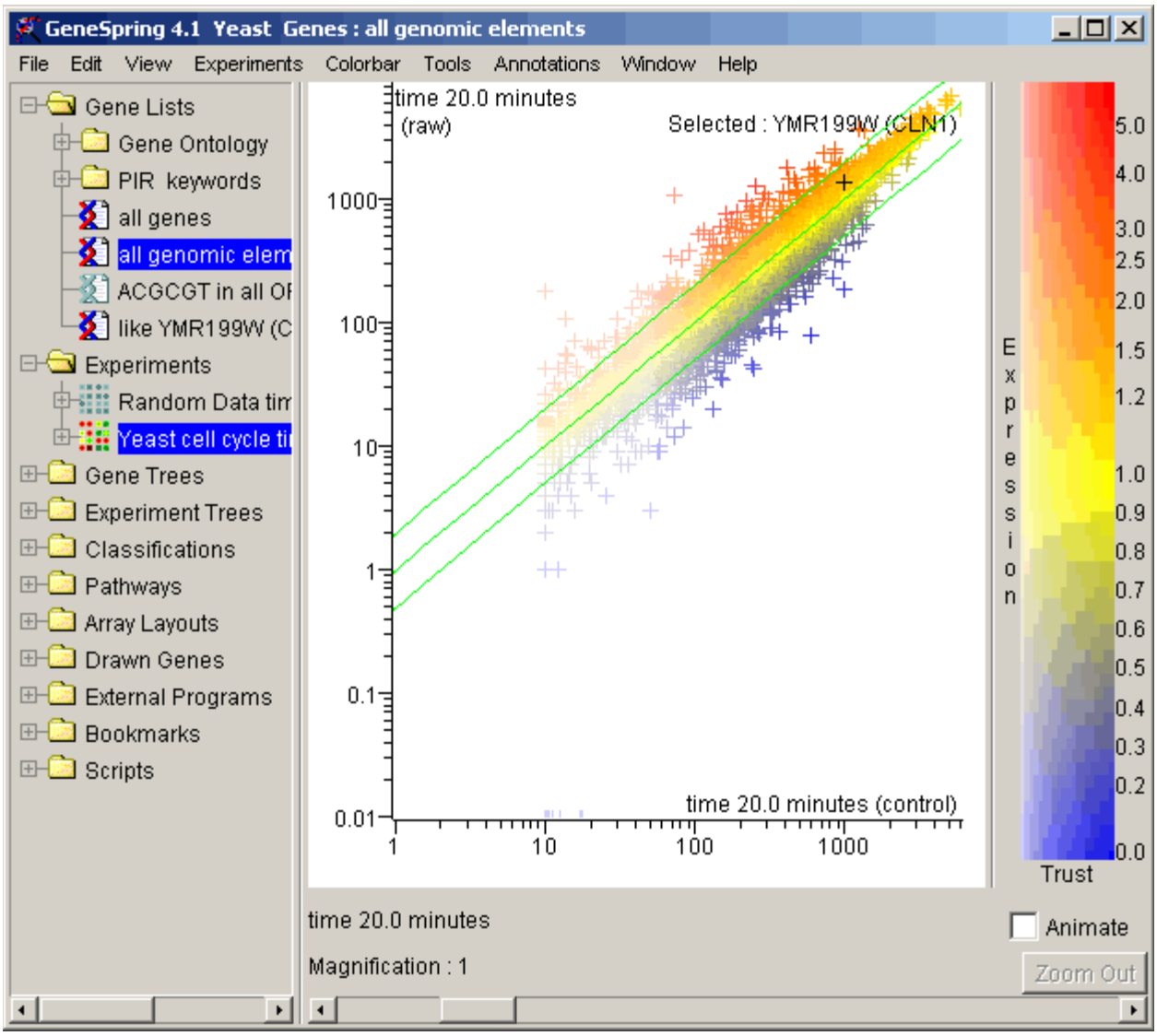


Figure 3-7 The Scatter Plot view

In the scatter plot in Figure 3-7, each '+' symbol represents a gene. The vertical position of each gene represents its expression level in the current condition, and the horizontal position represents its control strength (in this case, the median expression level of this gene in all conditions). Thus,

Copyright 1998-2001 Silicon Genetics

genes that fall above the diagonal are overexpressed and genes that fall below the diagonal are underexpressed as compared to their median expression level over the course of the experiment.

To view a Scatter Plot

1. Select the **`View > Scatter Plot`** option.
2. From the navigator panel, right-click the sample, condition, or gene list that you would like represented on the vertical axis and select the **`Use on Scatter Plot > Vertical Axis`** option from the drop-down menu.
3. From the navigator panel, right-click the sample, condition, or gene list that you would like represented on the horizontal axis and select the **`Use on Scatter Plot > Horizontal Axis`** option from the drop-down menu.
4. Right-click the horizontal axis and select the **`Horizontal Axis Mode`** option. Select one of the following data types from the submenu that appears:
   - **Relative (normalized):** to display the normalized expression value as defined in the current experiment (this is the most common option).
   - **Control:** to display the control signal as defined in the current experiment. See “Per-chip Normalizations” on page 2-22.
   - **Raw Signal:** to display the raw signal without normalizations applied.
   - **Average of Raw and Control:** to display the mean of the raw and control signals.
   - **Max of Raw and Control:** to display the higher of the raw or control signal.
5. Right-click the vertical axis, select the **`Vertical Axis Mode`** submenu, and choose an option as in step 4.
6. You can further modify the appearance of the plot by right-clicking the genome browser and selecting one of the following from the **`Options`** submenu.
   - **Show Lines or Hide Lines:** to add or remove the diagonal fold-lines.
   - **Use Big Points or Use Small Points:** to change the size of the symbols that represent genes.
   - **Show Gene Names or Hide Gene Names:** to show or hide gene names that appear beside the genes.

Copyright 1998-2001 Silicon Genetics

Tree View

The Tree view allows you to visualize your experiment as a mock phylogenetic tree, or dendrogram. In a tree, genes having similar expression patterns are clustered together.

1. From the navigator, open the Gene Trees or the Experiment Trees folder.
2. Click a tree name to select it. If there are no trees available for viewing you will need to create one.

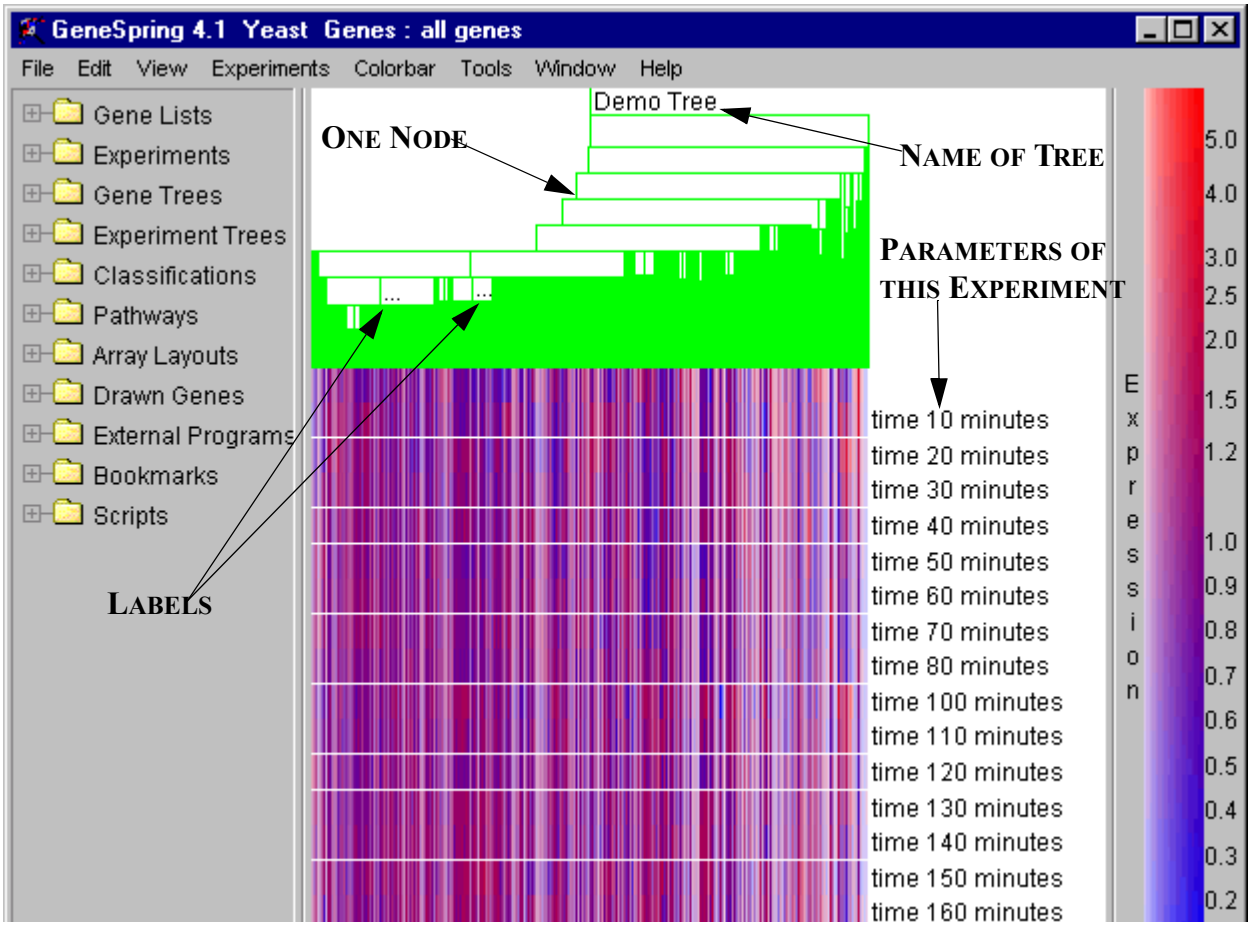


Figure 3-8 Tree View with annotations

The genome browser in Figure 3-8 is displaying a gene tree. The genes are the colored rectangles down at the bottom, joined to each other by green lines. As there are over six thousand vertical green lines in this view of the yeast genome, they tend to blur into each other, producing a solid green bar. Similarly colored genes tend to be clustered together, as expected. This will hold true for different points in the experiment. You can see the color changes vertically, as the current continuous parameter is arranged down the right side.

Copyright 1998-2001 Silicon Genetics

Magnifying Trees

The magnification in the Tree View is not quite the same as in the other views due to the need to keep the genes in the view along with the immediate tree branches. The amount of magnification will be visible in the parameter specification area just below the genome browser.

Selecting and Viewing Subtrees

1. **`Zoom in`** as described in *GeneSpring Basics Instructional Manual* 6.1.3 "Zooming in on the Tree View" on page 6-7.
2. Select any node by clicking over its intersection with your cursor. All the genes associated with that node will change to your *selected* color.

A single green line ending in a gene is a branch of the gene tree. Each bar crossing a set of branches forms a node of the intersecting branches. The distance from gene X to the node connecting it to gene Y indicates how closely the genes X and Y are correlated. The shorter the distance, the higher the correlation is.

You can also create a new tree from a node of a larger tree. Select a node as described above, then right-click in the genome browser and select **`Make Subtree`** from the pop-up menu.

Viewing Nodes

After clustering the genes according to their expression patterns, GeneSpring checks all known lists against all subtrees of the new gene tree, to assign names to the tree nodes where possible. These labels are taken from the gene lists in the standard lists.

- Place your cursor as close as possible to a label or intersection to view the text. When the cursor pauses over an intersection, a label will appear. It will disappear when the cursor is moved.
  All of the branches intersecting to form a node constitute the subtree defined by that node. A label such as "ribosome [15.1]" means the subtree from that node has a lot in common with the genes in the "ribosome" list. The numbers in square brackets are a measure of statistical significance. The higher the value, the more significant the comparison is. The comparisons between the lists and the subtrees are not looking for exact matches, but rather statistically significant overlaps, which may include subsets and supersets. When there is enough space on the screen, a label, if one exists, will be displayed along the top (horizontal bar) of the subtree. Otherwise, when there is space, a "..." will be displayed. An "&" symbol after a list name indicates the subtree is statistically similar to more than one list, all of whom, when there is enough room, are displayed as labels along the top of the subtree.

If you want to take a screen shot that includes the label, hover your cursor over the node, take the screen shot when the label appears. For most Windows applications, the cursor will not be visible, just the label. For more information about screen shots, please refer to "Saving Pictures and Printing" on page 6-2.

Copyright 1998-2001 Silicon Genetics

Viewing Gene Names in Trees

You can magnify the tree until the names are visible along the edge of the genes.

1. Place your cursor anywhere over the group of genes to view the gene name. When the cursor pauses over a gene, a label will appear. It will disappear when the cursor is moved.
2. Click once and that gene will become the selected gene. The name of the selected gene will appear in the upper right corner of the genome browser.

Viewing Colors in Trees

The coloring scheme of the current view is shown in the colorbar on the right. You can change the colors to any of the standard coloration options.

Color by all Conditions/Color by a Single Condition

In the **`Color by a Single Condition`** option, the genes in the gene tree are colored according to their expression at the condition indicated by the scroll bar at the bottom. With the **`Color by all Conditions`** option the genes in the gene tree are colored corresponding to each condition in the experiment, as shown by the name of the continuous parameter displayed at the right of the screen. The beginning of the experiment is colored at the top of the gene, next to the green line, and proceeds chronologically downward.

To Color by all Conditions:

1. Right-click while the cursor is in the genome browser.
2. A menu will appear, go to the **`Options`** submenu
3. Select **`Color by all Conditions`**.

To Color by a Single Condition:

1. Right-click while the cursor is in the genome browser.
2. A menu will appear, go to the **`Options`** submenu
3. Select **`Color By Single Condition`**.

Once your experiment is colored by single conditions, you can use the animate feature:

1. Select the Animate checkbox (a check mark will appear in the box when selected).

   Or,

1. Move the slider along the bottom of the main GeneSpring screen.

It may take a second or so for the tree to redraw when the time changes, because of the complexity of the picture.

Viewing Parameters in Trees

For most experiments, each measurement was taken under certain conditions. These conditions are listed in the far right side of the tree view. If one of the parameters has been designated as a

Copyright 1998-2001 Silicon Genetics

continuous parameter, it will be shown directly beneath the genome browser. The continuous parameter can be viewed with the animate command, if you first change the coloration to a single condition.

1. Right-click in the genome browser.
2. Select **Options > Color by a Single Condition**.
3. Select the **Animate** checkbox or use the slider at the bottom of the screen to change the condition displayed.

Horizontal Genes/Vertical Genes

It is possible to change the orientation of your Gene or Experiment Tree.

1. Right-click in the genome browser, and select **Options > Vertical Genes**.

Copyright 1998-2001 Silicon Genetics

Ordered List View

Allows you to view a gene list in the order of its associated values. Values are listed in descending order. If you do not have associated values, genes will be ordered according to the way they are listed in the Master Gene Table. Vertical lines representing genes are proportional to the gene's associated number.

To view genes in an ordered list, go to **`View > Ordered List`**. Your list will appear in its order.

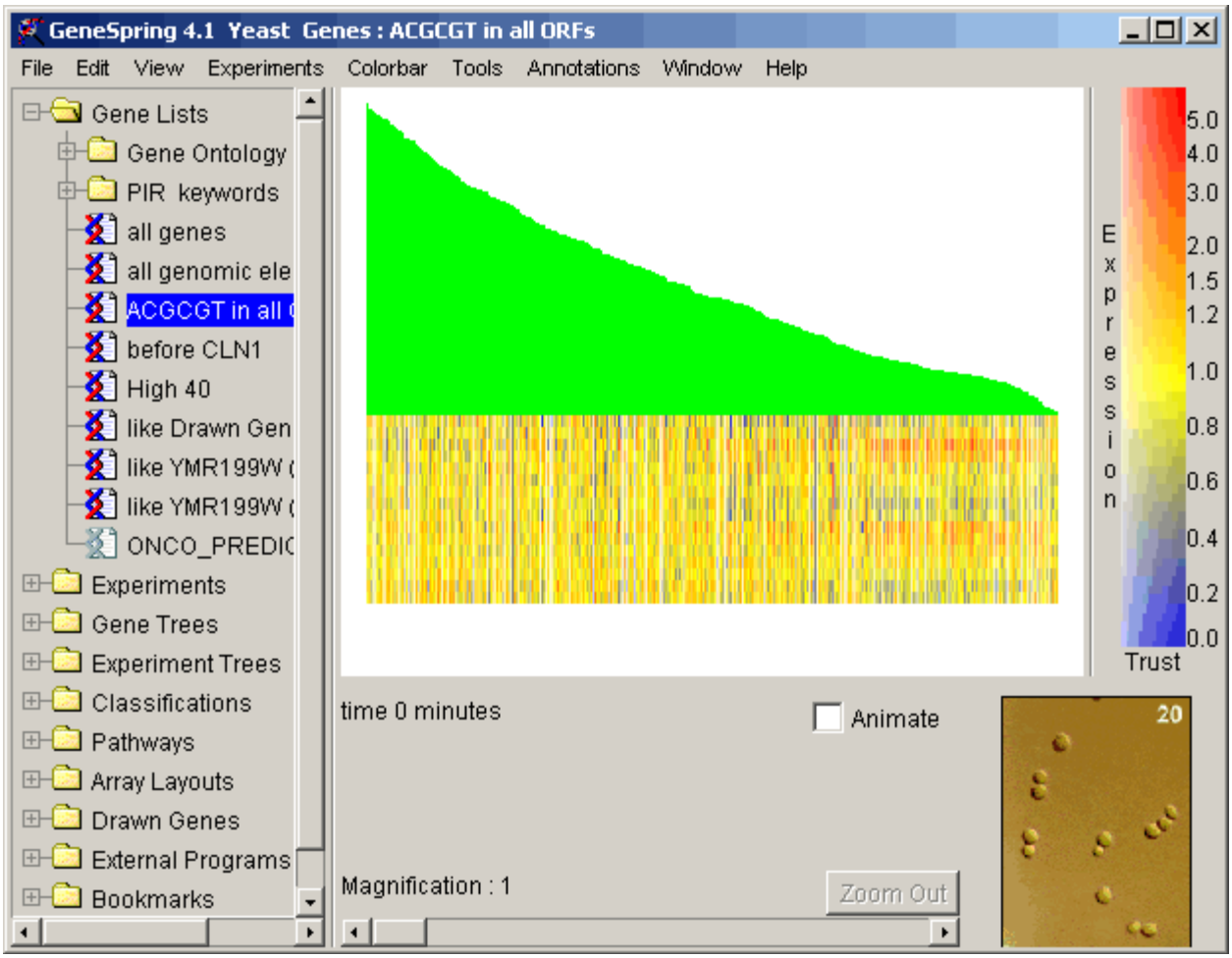


Figure 3-9 Ordered List View

To reach the following commands, right-click in the genome browser and select the **`Options`** drop-down menu.

- **Color by Single Condition/Color by All Conditions**—Allows you to visualize your data one condition at a time, where the slider dictates the condition (as in the Graph view), or to visualize all conditions at once, where conditions are layered one on top of the other, and the slider has no relevance.
- **Show/Hide Associated Values**—Shows or hides your associated values.

Copyright 1998-2001 Silicon Genetics

Array Layout View

The Array Layout view produces a synthetic picture of the arrays used in the current experiment. This view is useful in identifying arrays that display local shifts in intensity due to problems in probe deposition, hybridization, washing, or blocking. To use this view you must first create an array layout file (see “Creating an Array in GeneSpring” on page M-1).

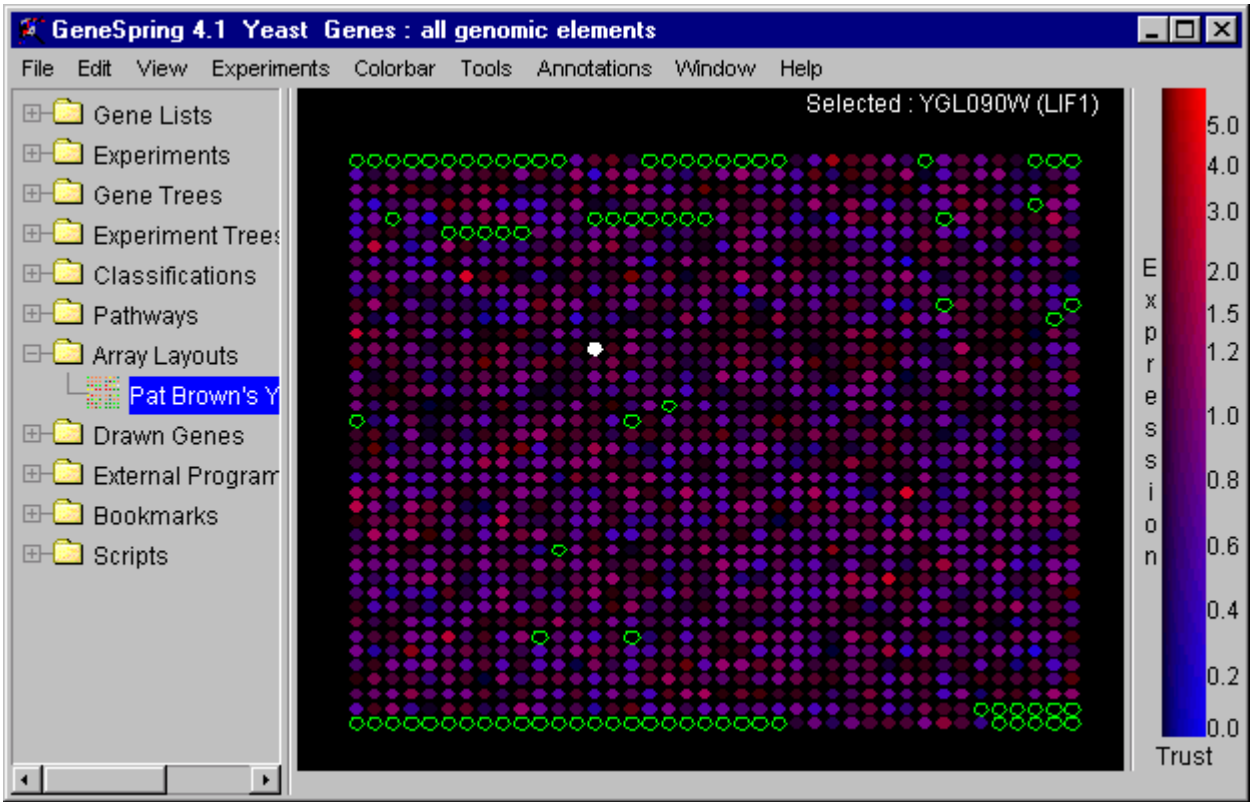


Figure 3-10 The Array view

In Figure 3-10 each solid circle represents an oligonucleotide on the array. If you zoom in, the gene names will become visible.

To view an Array Layout

1. Select the **`View > Array Layout`** option.
2. Select an array from the navigator.

Copyright 1998-2001 Silicon Genetics

Pathway View

The Pathway view lets you display and place genes on an imported *.gif* or *.jpeg* image.

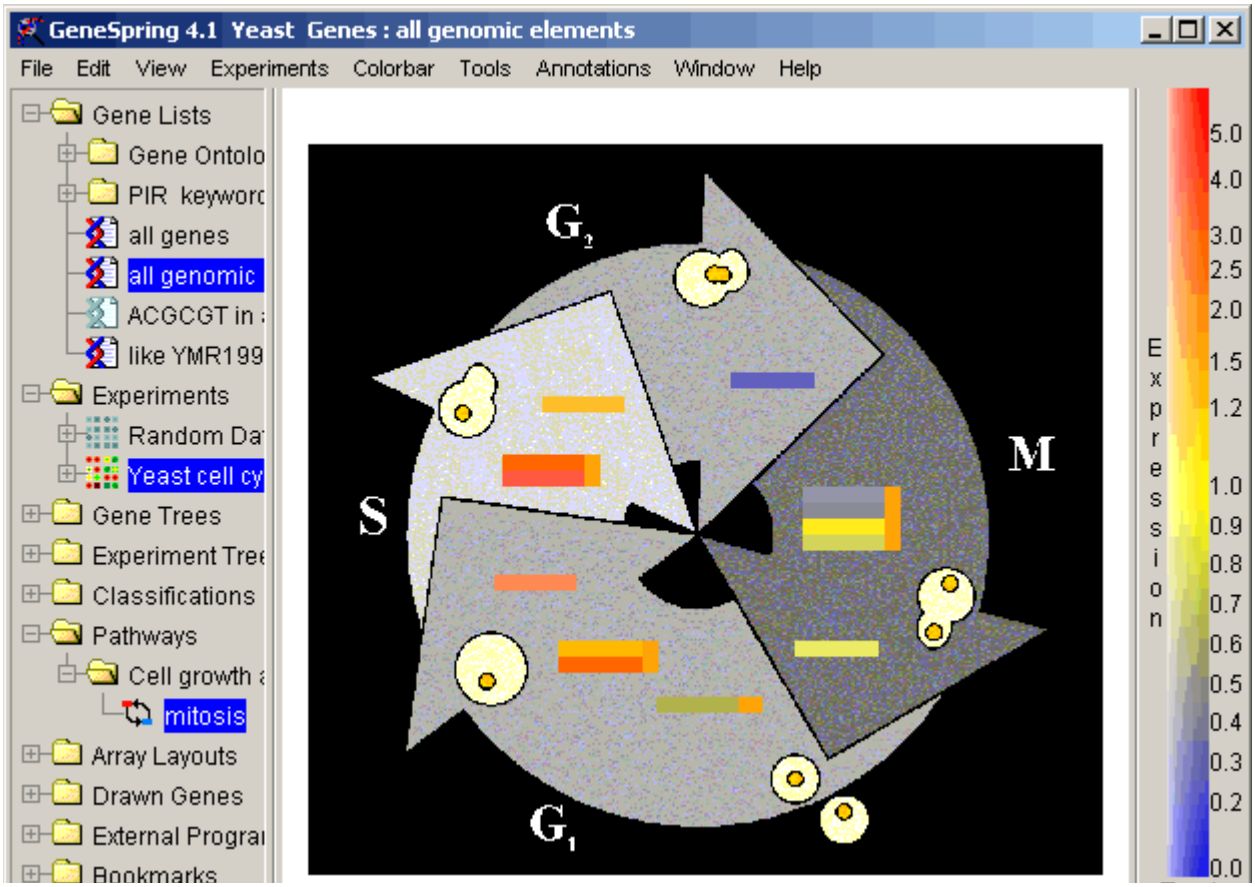


Figure 3-11 The Pathway view

To view a Pathway

1. Select a pathway from the Pathways folder in the navigator. (You will need to have already created a Pathway. See "Pathways" on page 4-23.)
2. Select a gene list. If a pathway contains a gene on a selected gene list, then the gene will be colored according to its expression level.

See the example of the mitosis pathway in Figure 3-11.

- To add a gene to the pathway, hold **`Ctrl`** and drag mouse over the desired placement area. Type a gene name or keyword. If a keyword is used, select the gene from the resulting list.
- To delete a gene from the pathway, right-click over the gene and select **`Delete Pathway Element`**.

Zooming, coloration, movement and the **`Find Genes Which Could Fit Here`** features work in this view. Find Genes Which Could Fit Here suggests genes that might be appropriate in certain areas of the picture. Please refer to the Pathways chapter for more details.

Copyright 1998-2001 Silicon Genetics

Compare Genes to Genes

The Compare Genes to Genes view allows you to observe the similarity between the expression profiles of two genes in one list or in two separate lists. Genes being compared are listed along respective graph axes. The correlation between any two genes is shown by a colored square at their point of intersection. Strong correlations in expression level are shown by a higher intensity color, weak correlations by a lower intensity color.

Associated values for gene lists are shown as lines extending perpendicularly from each axis. The length of the line represents the magnitude of the associated value. You can view these associated values by zooming in on the ends of the lines.

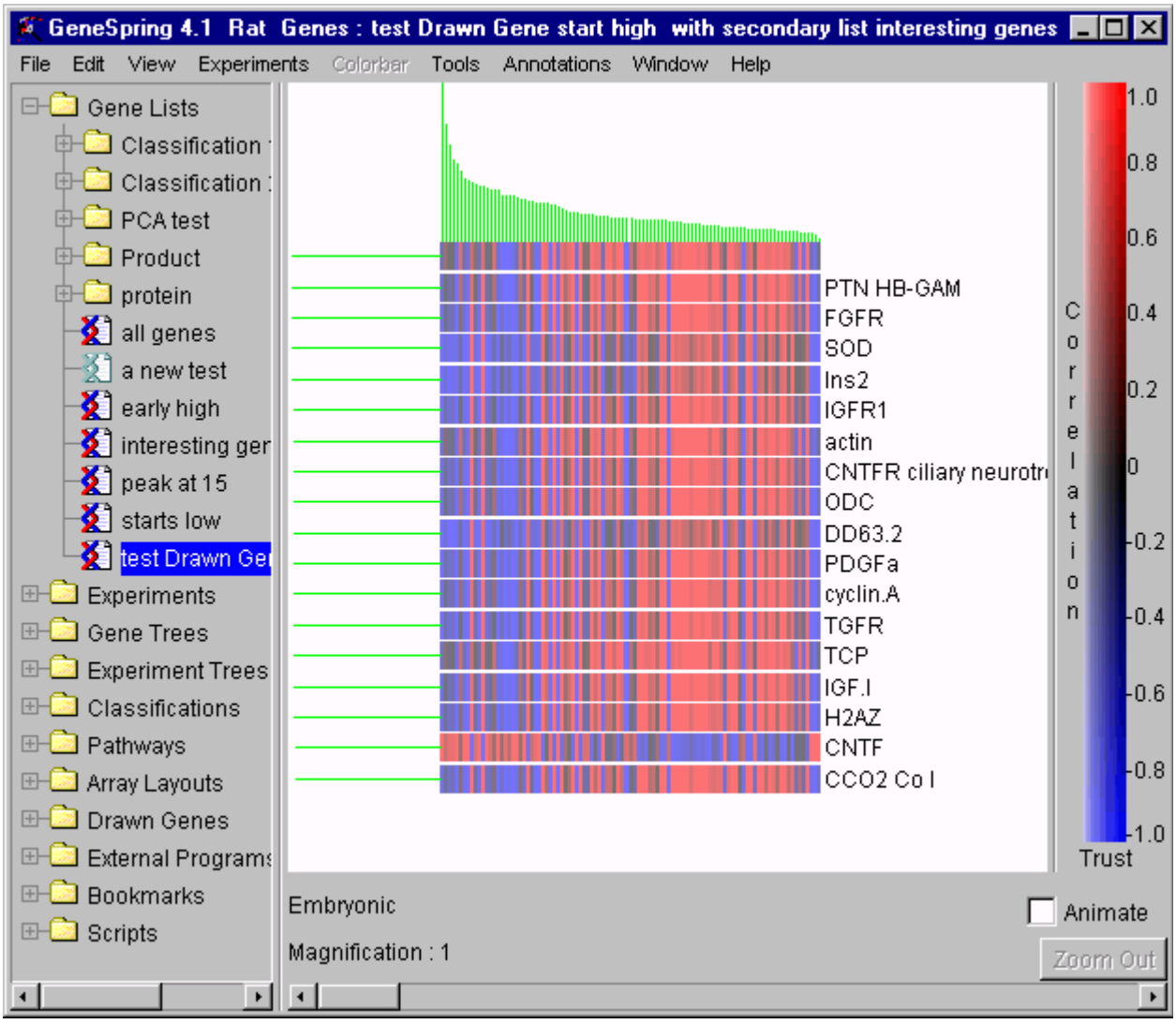


Figure 3-12 Compare Genes to Genes

Copyright 1998-2001 Silicon Genetics

In the Compare Genes to Genes view, GeneSpring employs a Pearson correlation to measure the pair-wise similarities (see “Pearson Correlation” on page L-2). Note that if you place the same list on both axes, you will see a line of perfect correlation values descending diagonally across the grid.

To view Compare Genes to Genes

1. Click the first gene list that you wish to compare in the navigator. (Please do this before you switch the view type, as large gene lists will take a very long time to compare.)
2. Select the **`View > Compare Genes to Genes`** option. The default display will place the selected gene list on both axes.
3. If desired, select a second gene list from the navigator by right-clicking on a gene list and selecting the **`Display as second list option`**. To remove this second list, select the **`View > Remove Secondary Gene List`**.

Copyright 1998-2001 Silicon Genetics

Graph by Genes View

The Graph by Genes view allows you to visualize an experiment as one line, where each point on the line represents the relative expression of one gene.

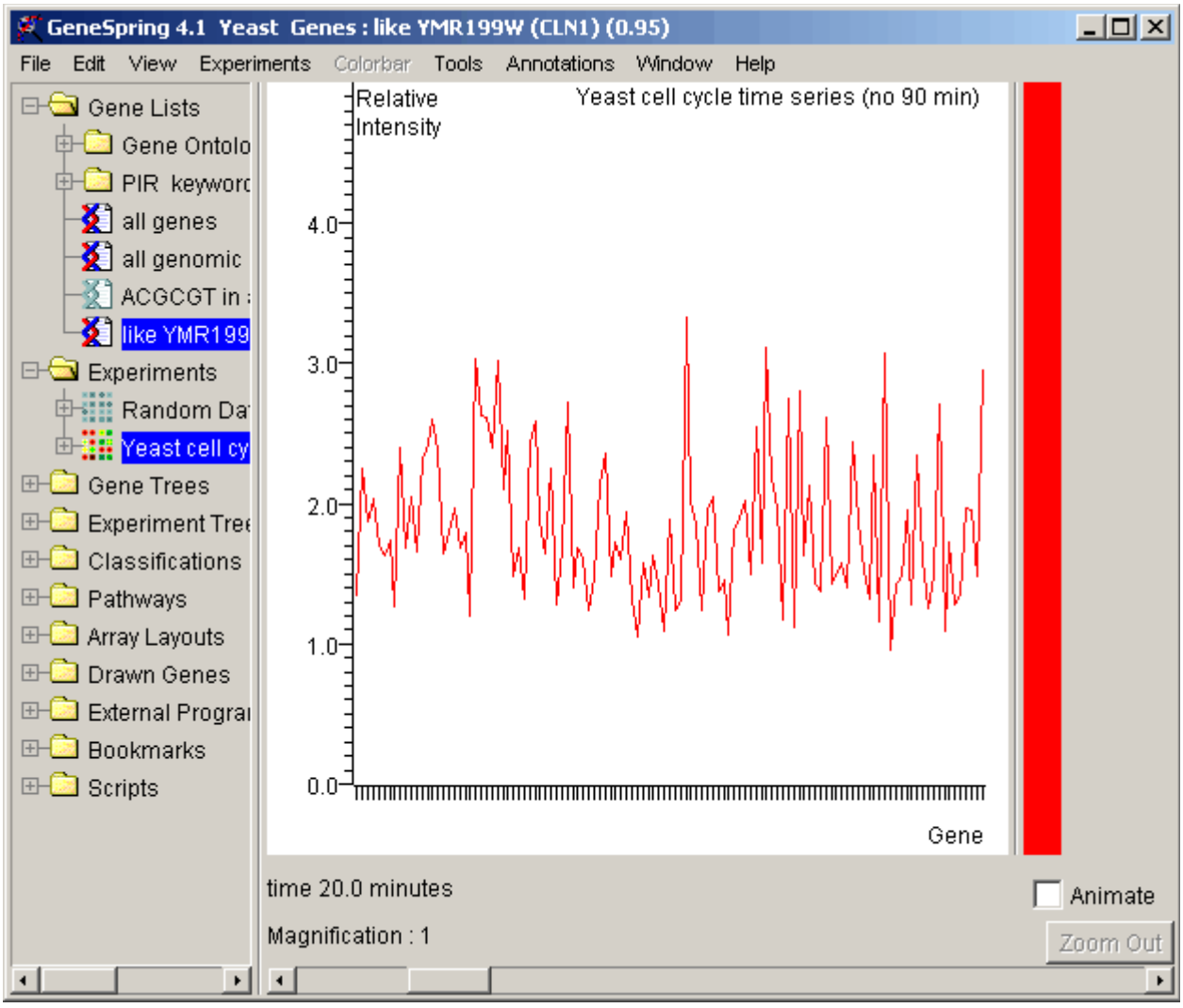


Figure 3-13 The Graph by Genes view, limited to the "Like CLN1" list

Figure 3-13 shows the genes in the "like YMR199W(CLN1)(0.95)" list in Graph by Genes view.

Genes at the top of the selected gene list are displayed at the left end of the experiment line and genes at the bottom of the gene list are displayed at the right end of the experiment line. Generally, your gene lists will be ordered so that the associated values appear in descending order. If you do not have associated values, your genes will appear in the same order as in the Master Gene Table.

To select a gene in the Graph by Genes view, you must use the **`Edit > Find Gene`** command. Clicking directly on the experiment line will not produce any results.

Copyright 1998-2001 Silicon Genetics

Functional Classification

It is possible to display genes according to some classification system. The Classification View is the usual way to display unsequenced organisms. Generally, the classification can come from either proprietary data which has assigned a label to each gene, or it can come from a set of lists, such as the Gene Onology lists already in the Gene Lists folder of the default yeast genome. You can also create classifications using GeneSpring's various features.

Coloring According to a Folder of Lists

As an example, these are the instructions to create a classification view with the Gene Ontology Lists.

1. Select **`View > Classification`**. You will see an unsorted classification.
2. Open the Gene Lists folder in the navigator. Open the Gene Ontology subfolder. Position the cursor over the **`biological process lists`** subfolder and click the right button, getting a pop-up menu. The command, **`Use as Classification`** will be at the top.
3. Select **`Use as Classification`** option. This makes the gene lists in the selected folder the classifications for the genes being displayed. The result should look like several lines of genes across the genome browser.
4. Zoom in. If your computer screen is small you may not be able to see the classification names and you will need to enlarge GeneSpring's main screen. Make the screen bigger by clicking the border and dragging the borders outwards. In particular, make the screen taller. You can also click and drag at the edges of the genome browser, making the navigator and the colorbar smaller.

Copyright 1998-2001 Silicon Genetics

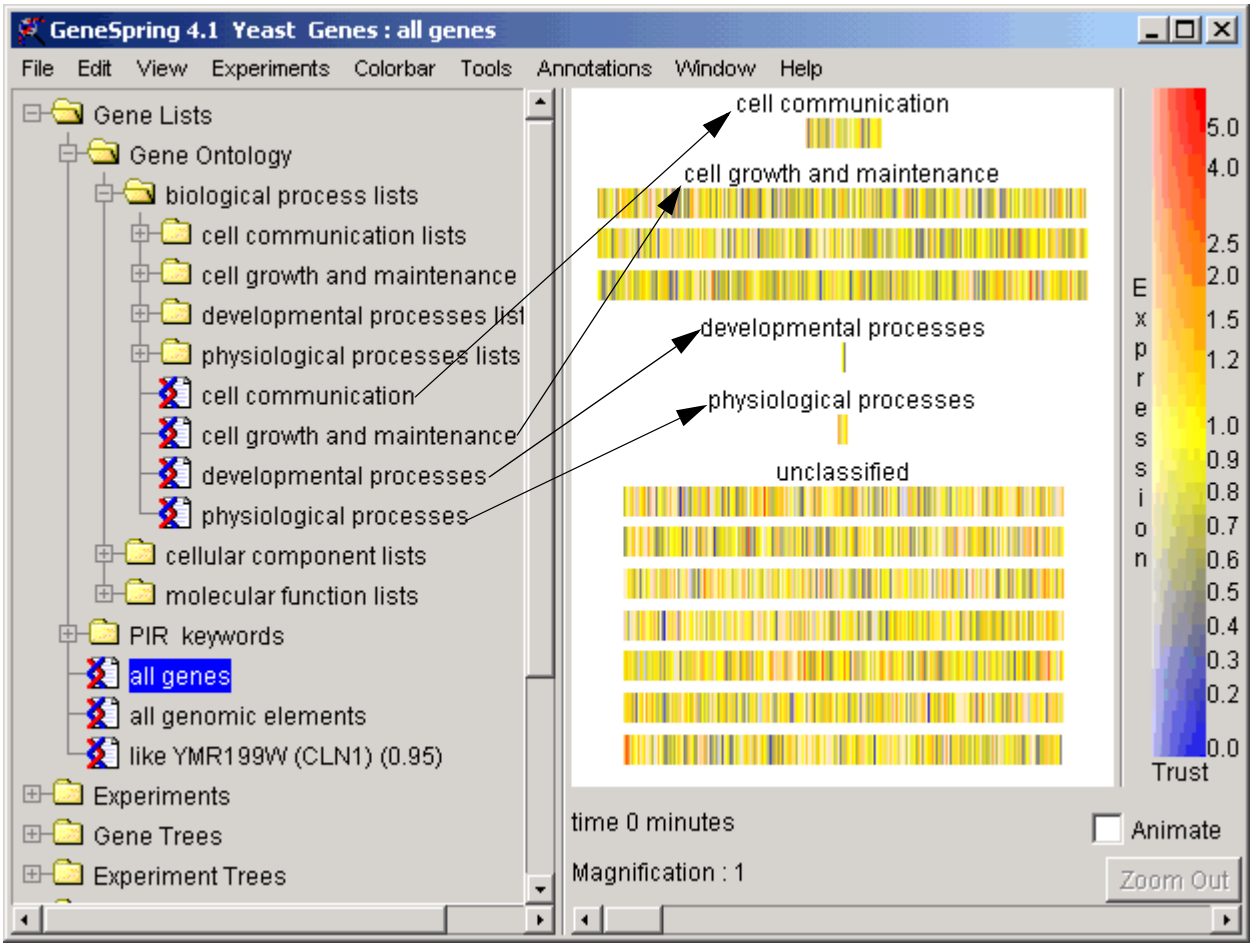


Figure 3-14 The Classification View

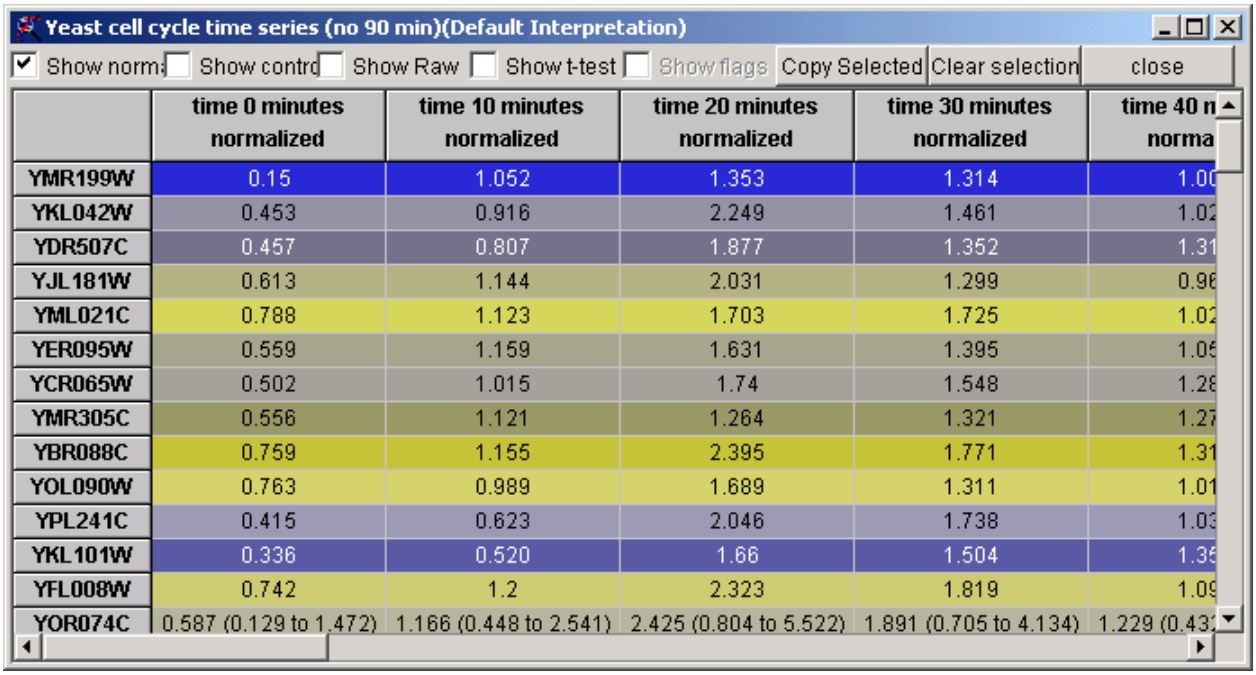
Each gene is divided up according to the gene lists in the Gene Onology Function subfolder, with the genes listed below their classifications. It is not surprising, given the source of the classification, that there are many “cell growth and maintenance” genes. You could choose any other gene list to view by selecting it in the navigator.

Once fully zoomed in, you can easily see the individual genes as small, distinct rectangles. You can zoom in to see some genes in greater detail. The gene names and the sequence will appear when there is enough space. It is possible for a single gene to be in more than one group, in which case it will be displayed in the first vertical group it is in. Genes not mentioned in any of the gene lists end up in the “unclassified” section on the bottom. The “unclassified” classification is a list of genes actively specified as unclassified. Some classifications may contain no genes depending on the list you are currently viewing. To clear a classification (and return the genome browser to the unsorted state) right-click over the Classifications folder in the navigator and select **`Clear Classification`** from the pop-up window.

Copyright 1998-2001 Silicon Genetics

View as Spreadsheet

Allows you to view your data as a spreadsheet. The spreadsheet color scheme and gene list reflect what is showing in the genome browser at the time you activate the new window. The order of the genes is the same as in your Master Table of Genes.



	time 0 minutes normalized	time 10 minutes normalized	time 20 minutes normalized	time 30 minutes normalized	time 40 n norma
YMR199W	0.15	1.052	1.353	1.314	1.0
YKL042W	0.453	0.916	2.249	1.461	1.02
YDR507C	0.457	0.807	1.877	1.352	1.31
YJL181W	0.613	1.144	2.031	1.299	0.98
YML021C	0.788	1.123	1.703	1.725	1.02
YER095W	0.559	1.159	1.631	1.395	1.05
YCR065W	0.502	1.015	1.74	1.548	1.28
YMR305C	0.556	1.121	1.264	1.321	1.27
YBR088C	0.759	1.155	2.395	1.771	1.31
YOL090W	0.763	0.989	1.689	1.311	1.01
YPL241C	0.415	0.623	2.046	1.738	1.03
YKL101W	0.336	0.520	1.66	1.504	1.35
YFL008W	0.742	1.2	2.323	1.819	1.09
YOR074C	0.587 (0.129 to 1.472)	1.166 (0.448 to 2.541)	2.425 (0.804 to 5.522)	1.891 (0.705 to 4.134)	1.229 (0.43

Figure 3-15 Spreadsheet View of the "Similar to CLN1" list

To Copy a Row for Pasting into another Document

1. Click on the row you wish to copy.
2. Right-click on the row and select **Copy**.

To copy the entire spreadsheet, click the **Copy All** button at the top right corner of the spreadsheet. Note that if you have any rows selected, you'll first have to click the **Clear Selection** button, also in the top right corner of the spreadsheet.

To Locate a Particular Gene

1. Type **Ctrl+F**.
2. Type in the gene name.
3. Click **OK**.

Inspect Found Gene

To bring up the Gene Inspector for your found gene, type **Ctrl+I**.

Copyright 1998-2001 Silicon Genetics

Linked Windows

Allows you to select one gene or gene list in two windows simultaneously. Simply select a gene or gene list in one window and the same gene or gene list will automatically be selected in the other window.

To create a linked window, go to the **File** menu and select **New Linked Window**.

Split Windows

Another interesting way to view classifications is with the Split windows function. The Split windows feature will allow you to see multiple sets simultaneously in the main GeneSpring screen.

To reach the split windows command, right-click over any item in the classification folder or any folder of classifications and move the cursor down to **Split window**. A small pop-menu will appear.

Select one of the options. If you selected **Vertically** the main screen of the genome browser will re-arrange into several small screens. Notice the number of genes in the upper right corner of each small screen.

While viewing split screens you can make changes in the experiment interpretation, zoom and pan around the same way you do with unsplit screens.

1. Right-click over folder **> Use as Classification**
2. Right-click **> Split window > Vertical**.
3. **View > Graph**.

You can double-click the banner bar to increase the screen size.

To unsplit the screen, select **View > Unsplit window** or right-click over the original data object and select **Split > Neither**.

You can also hide the labels appearing in the main GeneSpring screen.

All of the Hide and Show commands are simple toggle switches. Re-select that option to show what has been hidden. You may need to enlarge your screen before you can see all the labels.

Copyright 1998-2001 Silicon Genetics

Bookmarks

If you ever need to pause in the midst of your analysis, you can create a Bookmark to hold your place. The Bookmark saves all your current display settings, including experiment, gene list, coloration, and selected genes.

To Create a Bookmark

1. Go to the **File** menu and select **Save Bookmark**. The Save Bookmark dialog box will appear.
2. Name your bookmark.
3. Click **Save**.

To Access an Existing Bookmark

1. Click on the Bookmarks folder in the navigator.
2. Double-click over the name of any bookmark to open.

Or,

1. Go to **File** and select **Load Bookmark**. The Load Bookmark dialog box will appear.
2. Select your bookmark.
3. Click **Open**.

Changing the Coloring Scheme

Color by Expression

This option colors genes according to their normalized expression values and trustworthiness. To color your genes by expression, select **Colorbar > Color by Expression**.

Expression

The vertical axis of the colorbar represents expression levels on a continuous scale. Using the default colors, red indicates overexpression, yellow indicates average expression, and blue indicates underexpression. Genes are colored by their expression level in the selected condition as indicated by the condition line. If you have specified the parameter on the horizontal axis to be continuous, expression levels in between conditions will be interpolated.

Copyright 1998-2001 Silicon Genetics

Trust

The horizontal axis of the colorbar indicates the degree to which you can trust your data, where dark or unsaturated colors represent low trust, and bright, saturated colors represent high trust.

You can assign trust values for each gene when you load your experiment or allow GeneSpring to create trust values automatically (the latter numbers are listed in the Gene Inspector, in the *Control* column). To enter your own numbers, see "The Experiment Wizard" on page D-1. The following are the guidelines by which GeneSpring automatically creates trust values:

In **two-color experiments**, the trust value is usually the control channel (typically Cy5), unless you do a per-chip normalization in which case it is:

(the control channel) x (the median of the control channel) x (the median of the signal channel)

For **Affymetrix and other one-color experiments**, the trust value is constructed based on the normalizations you have chosen. If you accept the default normalizations for Affymetrix data (use distribution of all genes using the 50th percentile and normalize to the median for each gene), then trust is:

(the median value of the chip) x (the median value of the gene)

If you choose to use distribution of all genes using the 50th percentile and normalize to sample(s), trust is calculated as follows:

(the median value of the chip) x (the average of the gene's measurement in control samples)

GeneSpring automatically interprets trust for **Affymetrix data**, specifying 500 as data that is most trustworthy, 150 as moderately trustworthy, and 50 as least trustworthy. For other data, you will need to enter these numbers manually. Consult the manuals for your array scanning software or hardware and estimate these trust levels based on the detection limit and noise levels for any given measurement.

To set the trust interpretation:

1. Right-click the colorbar.
2. Click **`Set Range`**.
3. Enter values for High Control Strength, Medium Control Strength, and Low Control Strength.
4. Click **`OK`**.

Copyright 1998-2001 Silicon Genetics

Color by Significance

Data is colored based on how far the gene is over- or underexpressed (relative to a normalized expression level of 1), in terms of the standard error of the measurement. The standard colorbar is replaced with a colorbar ranging from +3σ to -3σ. The standard error model is based on the Global Error Model, if the Global Error Model is turned on. (For more information about the Global Error Model, see "Global Error Models" on page 2-26.) Otherwise the standard error is based on the standard deviation of the replicate data for a particular gene and condition (for information about the calculation of this error, see the Gene Inspector).

To color your genes by significance, select **`Colorbar > Color by Significance`**.

Color by Static Experiment

This option allows you to color your experiment by a single condition. The vertical axis of the colorbar represents relative intensity on a continuous scale. In the default coloration, red indicates overexpression, yellow indicates average expression, and blue indicates underexpression. The horizontal axis of the colorbar indicates the degree to which you can trust your data, where dark, or less intense, color represents low trust, and light, or more intense color, represents high trust (for information about trust, see "Trust" on page 3-32).

To Color by Static Experiment

1. Click the **+** sign to the left of your experiment in the navigator.
2. Click the **+** sign to the left of your experiment interpretation.
3. Right-click over the condition you wish to color by and select **`Set Static Experiment`**.

To deselect color by static experiment, go to the **`Colorbar`** menu and select a different coloring scheme.

Color by Venn Diagram

This option colors genes based on their membership in one or more gene lists in a Venn diagram. For information about creating Venn diagrams and using them for analysis, see "Making Lists with the Venn Diagram" on page 4-19.

Color by Parameter

This option colors genes based on the value of parameters. This coloring scheme is best suited for use with Graph view and Bar Graph view when different conditions are indicated with discrete symbols.

To Color by Parameter

1. Select **`Experiments > Change Experiment Interpretation`**.
2. Choose the parameter(s) you wish to color by and click **`Color Code`** for that parameter. Click **SAVE** to create a new interpretation.

Copyright 1998-2001 Silicon Genetics

3. Select **`Colorbar > Color by Parameter`**.

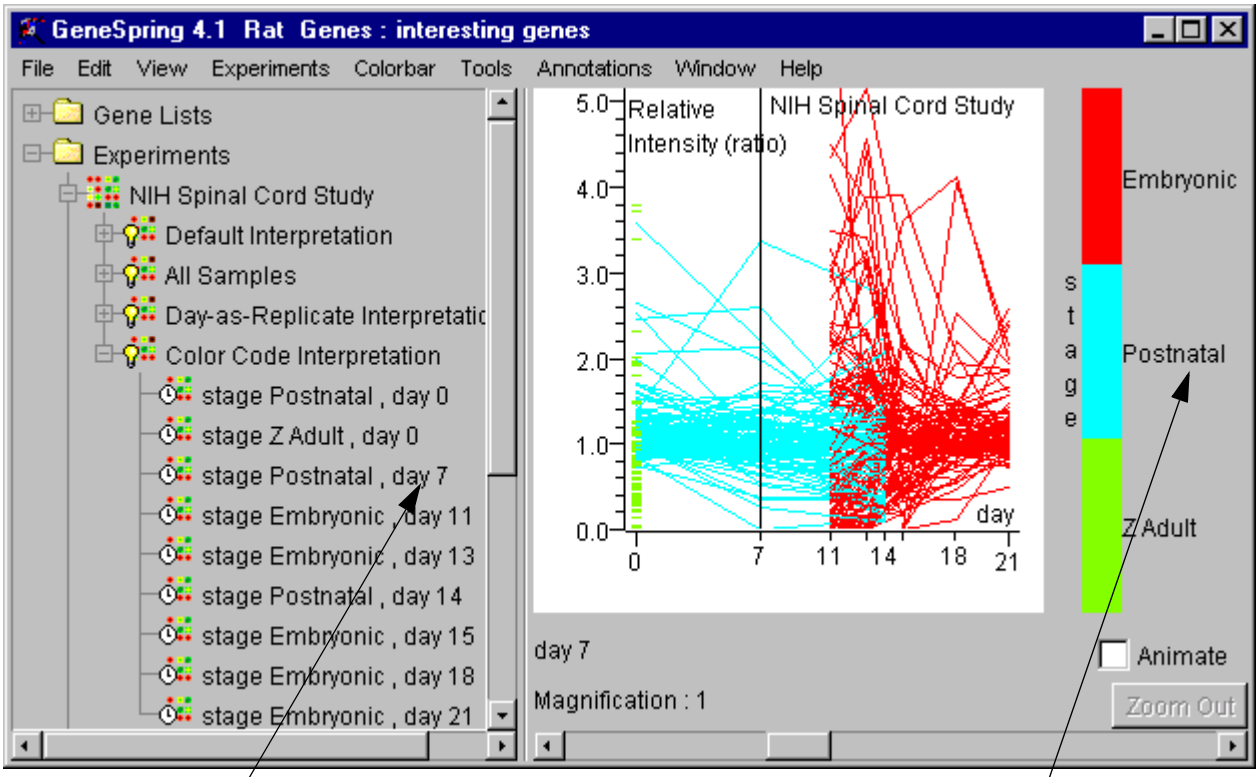


Figure 3-16 The NIH Spinal Cord Study colored by parameter

No Color

This option allows you to view genes with no coloration, showing all genes in gray.

To implement this option, select **`Colorbar > No Color`**.

Color by Classification

This coloring scheme allows you to color-code the genes by some previously defined knowledge about them. You can use a folder of lists to color by classification or a classification method such as k-means or SOM.

Coloring a Previously Saved Classification

You can use a previously saved classification for coloring.

1. Open the **`Classifications`** folder by clicking its icon.
2. Select a classification by right-clicking over the name.
3. Select **`Set as coloring scheme`** from the pop-up menu and GeneSpring will automatically update to reflect the new coloring scheme.

Copyright 1998-2001 Silicon Genetics

The colorbar will show the names of the sets present in the chosen classification.

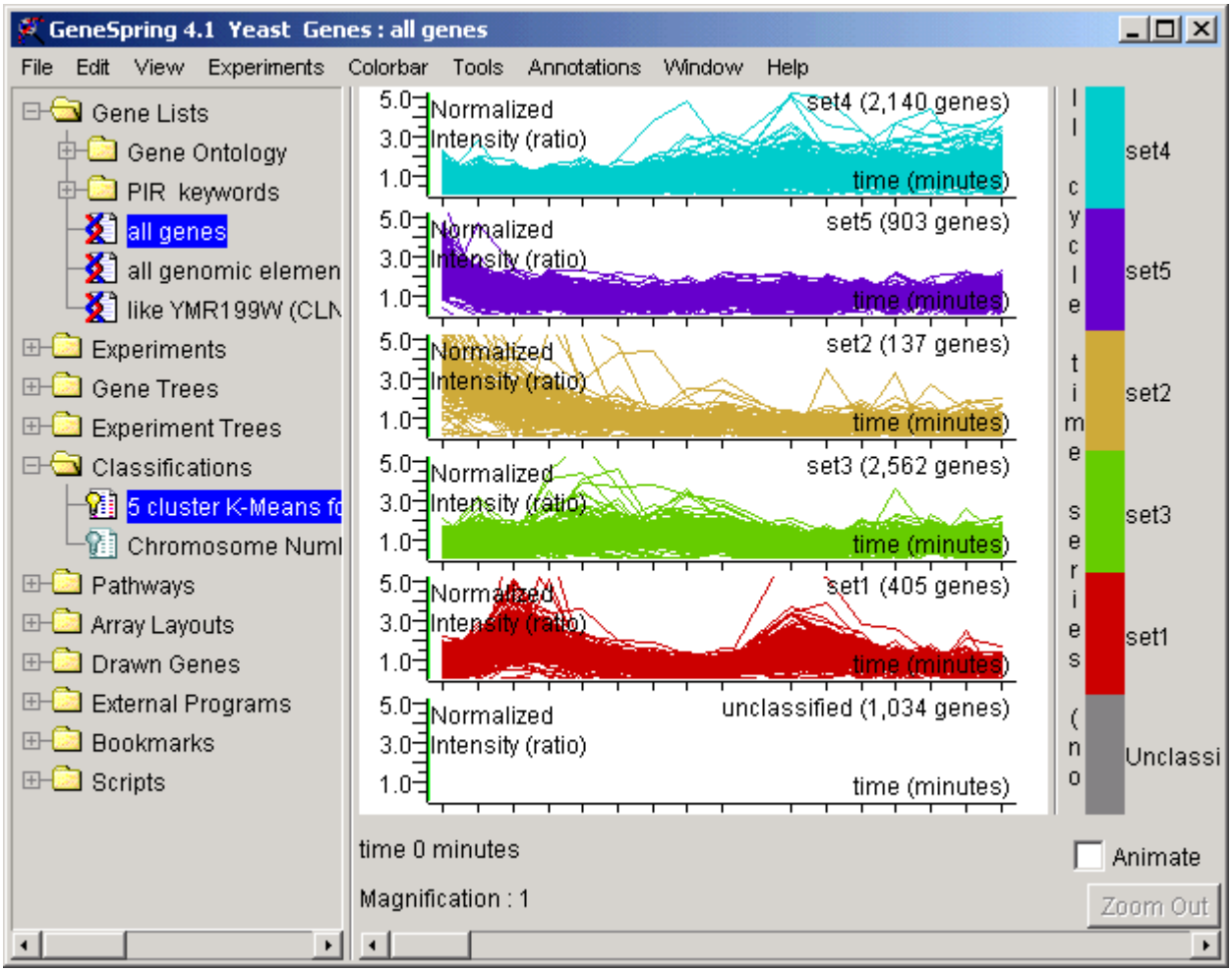


Figure 3-17 A Split Window, colored by Classification

Split Window and Color by Classification

You can also use the Split Window feature with the Color by Classification scheme.

1. Select a gene list to view.
2. Right-click over a folder or a previously saved classification and select **`Use as Classification`**.
3. Right-click over that folder again and select **`Split Window > Both`**.

Color by Secondary Experiment

The Graph and Scatter Plot displays lend themselves to being colored in many different ways because the display presents expression levels of the genes through the entire experiment. These are the only views in which you may choose to color the genes by a secondary experiment. This means the color of each gene line graphed correlates to the expression level of that gene in a different experiment, at the point in the second experiment marked by the secondary scroll bar.

Copyright 1998-2001 Silicon Genetics

1. From the navigator, open the **Experiments** folder by clicking on its icon.
2. Position your cursor over an experiment (not the one currently displayed) you would like to use for coloration.
3. Right-click and select **Set Secondary Experiment** from the pop-up menu.

The coloring scheme of the genome browser will be shown in the colorbar on the right. There will be two versions of the animation controls in the Experiment Specification Area.

Changing the Experimental Data Range

Before you change the experimental data range, you will need to select **Colorbar > Color by Expression**.

1. Right-click over the colorbar and select **Set Range** from the pop-up menu.
2. Reset the values determining the intensity of the colors used by the genome browser.
3. Click **OK**.

There are six categories you can change:

- **High Expression—**High expression refers to the normalized expression of your genes, it is the vertical axis of the color bar. The default for this is 6.0.
- **Normal Expression—**For most normalization procedures in GeneSpring the data will be normalized to 1.0. The default for this is 1.0.
- **Low Expression—**For most normalization procedures in GeneSpring the data will not have negative numbers. The default for this is 0.0.
- **High Control—**High control refers to the control strength of your genes. It is represented by the horizontal axis of the colorbar. The default for this is 200.0.
- **Medium Control—**The default for this is 100.0.
- **Low Control—**The default for this is 50.0.

For example, you could change the usual range of an experiment to high = 10, normal = 5 and low = -2 resulting in a very different color scheme once you click **OK**.

There is no **Edit > Undo** (**Ctrl+Z**) function for this type of change. If you want to return to your previous coloration scheme, you must re-open the Experiment Data Range pop-up window and type in your old values.

For more details on trust, please see “Trust” on page 3-32. For more details on normalization, please see “Normalizing Options” on page G-1.

Copyright 1998-2001 Silicon Genetics

Changing the Default Colors

You can change the colors used by GeneSpring to display the genes. This will not affect the interpretation of your data, although it might help you to make genes more visible on-screen or make it easier to print screen shots.

1. Select **Edit > Preferences**.
2. In the drop-down menu, select **Colors**.
3. Select the type of information whose color you would like to change and click the **Change** button.
4. Adjust the slider until the color you want is displayed in the preview window at the top of the Preferences window.
5. Click **OK**.

For more details about the other options in the Preferences window, please refer to “Preferences Window” on page B-1.

The Inspectors

GeneSpring’s Inspectors are a series of windows allowing you to view the current defaults and available details of any gene, condition, classification or experiment.

Gene Inspector

One of GeneSpring’s most flexible tools, the Gene Inspector allows you to look at all the data associated with a particular gene, see the lists that include your gene, make correlations, and link directly to Internet databases.

In the upper left corner of the Gene Inspector window is the name of the gene and an area for notes. The table in the upper right corner displays the normalized, control, and raw values, as well as the t-test p-value and flag for each measurement. In the center of the window is a browser showing a graph of the gene across all conditions. At the bottom of the window, from left to right, are correlation functions, lists containing your gene, and links to databases.

To reach the Gene Inspector window

Double-click on a gene (this may be easier after zooming in)

Or,

1. Select **Edit > Find Gene**.
2. Enter the name of your gene.
3. Press **Ctrl+I**.

Copyright 1998-2001 Silicon Genetics

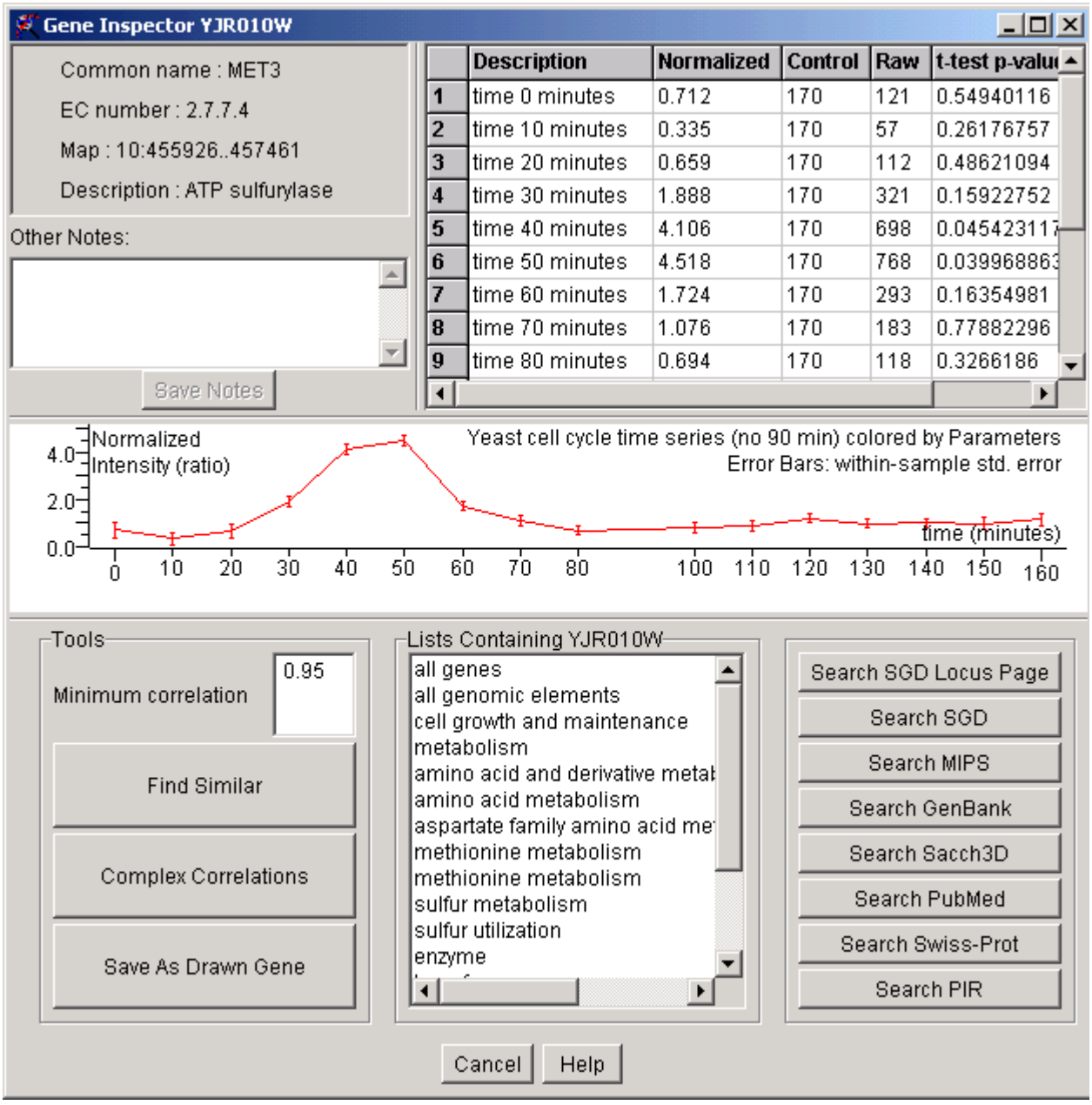


Figure 3-18 Gene Inspector window for gene MET3 (Yeast Cell Cycle)

Gene Identification Section

Information on the selected gene from the master gene table is displayed in the upper left corner of the Gene Inspector in the Gene Identification section.

Copyright 1998-2001 Silicon Genetics

The Data Table

The table in the upper right corner is the Data Table. It contains the following information:

- **Description—**The condition under which the measurement was taken.
- **Normalized—**The normalized data value. For information about normalizations. See "Experiment Normalizations" on page 2-21.
- **Control—**The control strength for the gene. For information about control strengths. See "Per-gene Normalizations" on page 2-25.
- **Raw—**The raw value of the data, just as it came off the chip or out of the scanner.
- **t-test p-value—**The t-test p-value is applicable only to replicated data. For information on this calculation, see "The T-test P-value" on page 3-39.
- **Flags—**Flags indicate whether or not your data is reliable. Whether or not you have flags will depend on your instrumentation and what you have entered into your master gene table. See "Measurement Flags" on page J-12.

The T-test P-value

In cases where there is replicate data, a one-sample Student's t-test is calculated to test whether the mean normalized expression level for the gene is statistically different from 1.0. The t-statistic is calculated as:

$$t = \frac{\overline{X} - 1}{S_X / \sqrt{n}},$$

where $\overline{X} = \sum_{i=1}^{n} X_i$ is the sample average of the n normalized expression levels $X_1, \ldots, X_n$, and

$$S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Figure 3-19 The formula for t-test

is the sample standard deviation of the replicates. The value of t is compared with a table of the distribution of Student's t-distribution with $n - 1$ degrees of freedom to yield the significance level (or *p-value*) for a two-sided test that the mean gene intensity differs significantly from 1.0.

The Browser Display

The Gene Inspector browser shows the gene's expression over the experimental parameter, time (minutes) in Figure 3-18. The browser image reflects the experiment interpretation in the main browser window. The only view option available in the Gene Inspector is the Graph view.

Copyright 1998-2001 Silicon Genetics

By right-clicking on the browser, you can use error bars in the browser display, create a resizable picture of the browser, or save a bookmark. By right-clicking and selecting **Options**, you can change the vertical axis range, show or hide many of the browser elements, and switch your view from normalized to raw data. For more information about the latter options, see "Using Genome Browser" on page 3-1. For information about error bars, see "Global Error Models" on page 2-26. For information about creating a resizable picture, see "Saving Pictures and Printing" on page 6-2. For information on bookmarks, see "Bookmarks" on page 3-31.

Gene Inspector Tools

The box in the bottom left corner of the Gene Inspector contains tools allowing you to search for genes having similar expression profiles to the gene being displayed in the Gene Inspector.

- **Find Similar—**Allows you to search for genes with similar expression profiles to the gene being inspected. Each gene expression profile must have the required minimum correlation to be considered similar. The higher the minimum correlation (maximum 1), the closer the gene expression profiles have to be. Enter this number in the Minimum correlation box above the Find Similar button. For information on using the Find Similar function, see "Making Lists with the Find Similar Command" on page 4-13.
- **Complex Correlation—**Allows you to make a gene list comparing the gene being inspected to genes having similar expression profiles in multiple experiments, with more complex parameters than the Find Similar tool allows. For information on using the Complex Correlation function, see "Making Lists with the Complex Correlation Command" on page 4-14.
- **Save As Drawn Gene—**Allows you to save your gene expression profile as a drawn gene, which you can use to make lists. For information on making lists from drawn genes, see "Creating Drawn Genes" on page 4-22.

Lists Containing Your Gene

In the bottom center of the Gene Inspector are the lists containing your gene. Selecting one of these lists will bring up the Inspect List window. For information about this window, see "List Inspector" on page 3-44.

Searching Internet Databases

You can set up the Gene Inspector to search public databases. To set-up this search function, see "Genome Wizard" on page C-1. Note, however, that the Macintosh version of GeneSpring does not allow for Gene Inspector searches of the Internet. To search a database with a Macintosh, go to **Edit > Preferences > Browser** and enter the appropriate pathway.

Notes Section

In the upper left corner of the Gene Inspector, under the Gene Identification Section, is an area where you can make notes. To save these notes, click the **Save Notes** button.

Copyright 1998-2001 Silicon Genetics

Experiment and Condition Inspectors

Just as you can inspect a gene with the Gene Inspector, you can inspect an experiment or conditions with the Experiment or Condition Inspector.

To Access the Experiment or Condition Inspectors

1. Right-click over the name of any experiment or condition in the navigator.
2. Select the **`Inspector`** option from the pop-up menu.

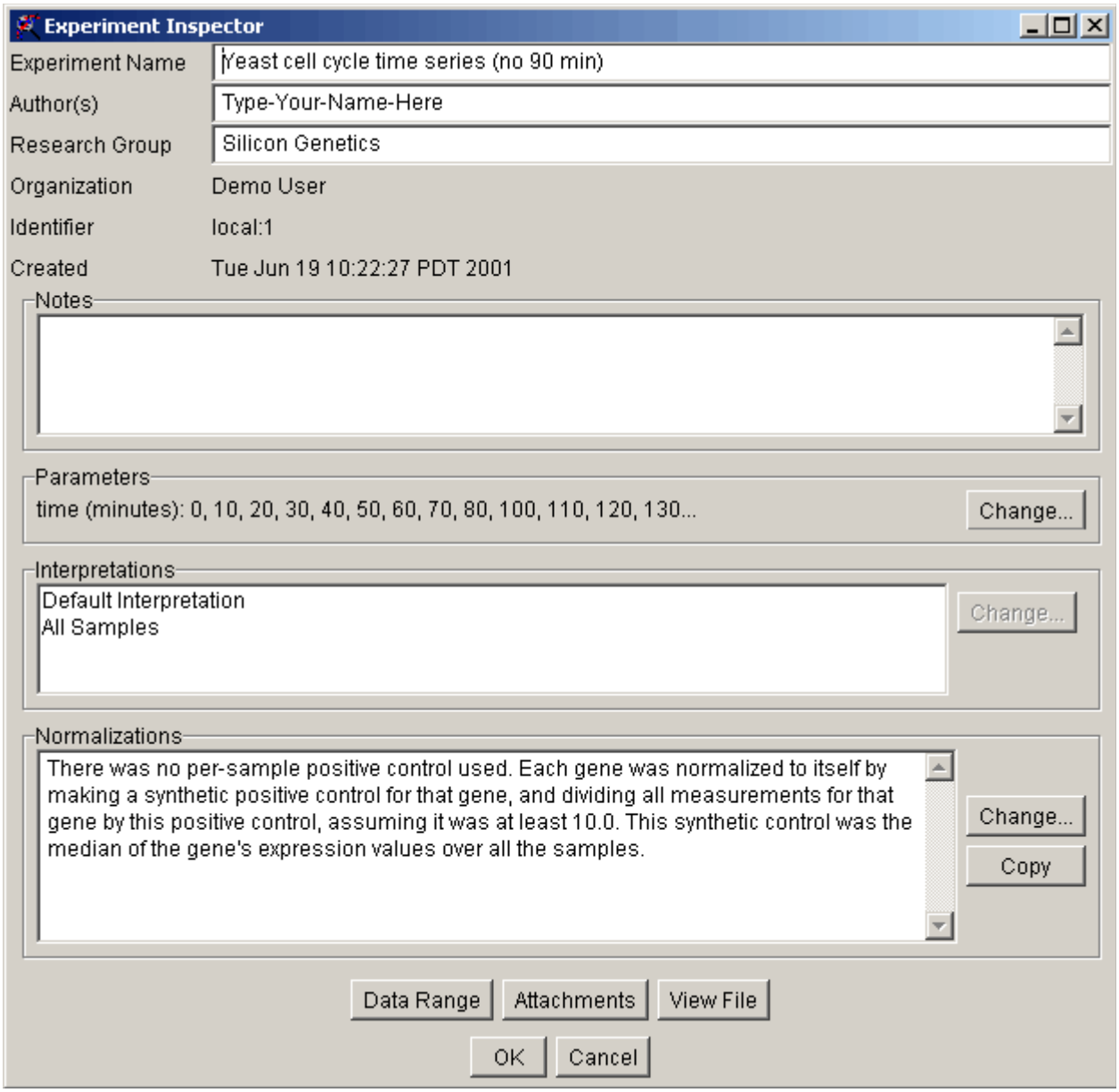


Figure 3-20 The Experiment Inspector window

Copyright 1998-2001 Silicon Genetics

The upper section of the Experiment Inspector contains the experiment information. Most of the text in the white boxes are directly editable. You can type, copy and paste as you do with any normal text editor.

The Parameters box

Within the parameters box you can view the various parameters for the experiment and their possible values.

Selecting the **Change** button in the parameters box will result in the Change Parameters window.

Please refer to “Change Experiment Parameters” on page 2-8 for details on this window. Any changes made in the Change Parameters window will be saved and affect your experiment when you click **OK**.

The Interpretations Box

A list of all the interpretations associated with this experiment is in the Interpretations section of the Experiment Inspector window. You can select any of the interpretations in the white text boxes by clicking over them. A double-click will bring up the Change Experiment Interpretation window automatically. If your computer is not set to acknowledge a double-click, select with a single click and select the **Change** button. This will bring up the Change Experiment Interpretation window. Please refer to “Changing the Experiment Interpretation” on page 2-17 for details on this window. Any changes made in this window will be saved and affect your experiment when you click **OK**.

The Normalizations Box

Near the bottom of the Experiment Inspector window is the Normalizations panel. Here, you can read what normalizations are currently being used in your experiment. If you would like to use the text elsewhere, you can click the **Copy** button and the text will be placed on your clipboard for use in other applications.

Selecting the **Change** button in the Normalizations box will result in the Experiment normalizations window.

Please refer to “Experiment Normalizations” on page 2-21 for details on this window. Any changes made in this window will be saved and affect your experiment when you click **OK**.

Copyright 1998-2001 Silicon Genetics

The Bottom Buttons

Across the bottom of the Experiment Inspector are several buttons.

- **Data Range—**The **`Data Range`** button will bring up the Data Range window. You can use this window to alter what measurements are considered *high*, *normal* or *low* in GeneSpring's coloration scheme. Any changes made in this window will be saved and affect your experiment when you click **`OK`**. For more information about the Data Range and how it affects the color your experiment is presented in the main GeneSpring window, please refer to "Changing the Experimental Data Range" on page 3-36
- **Attachments—**The **`Attachments`** button brings up an Attachments window.You can add any number of files or folders to your experiment from this window. Any changes made in this window will be saved when you click **`Close`**.
- **View File—**The **`View File`** button will launch your default browser and allow you to view all of the information associated with your experiment in HTML format.
- **OK—**This button will save all your data.
- **Cancel—**This button will close the Experiment Inspector window without saving any of the changes you may have made in any of the white text boxes.

Condition Inspector

A condition a unique combination of parameters as applied to your sample. Each condition may be a single sample or a group of replicate samples combined based upon the parameter values defined for each sample. The easiest way to think of this is as the parameters under which the sample(s) was observed. If you have no replicates, condition and sample can be considered synonymous.

1. Open the **`Experiment`** folder in the navigator by clicking on its icon.
2. Click the + sign next to the experiment icon.
3. Click the + sign next to the interpretation icon.
4. Right-click over a condition.
5. Select **`Inspect`** from the pop-up menu.

Copyright 1998-2001 Silicon Genetics

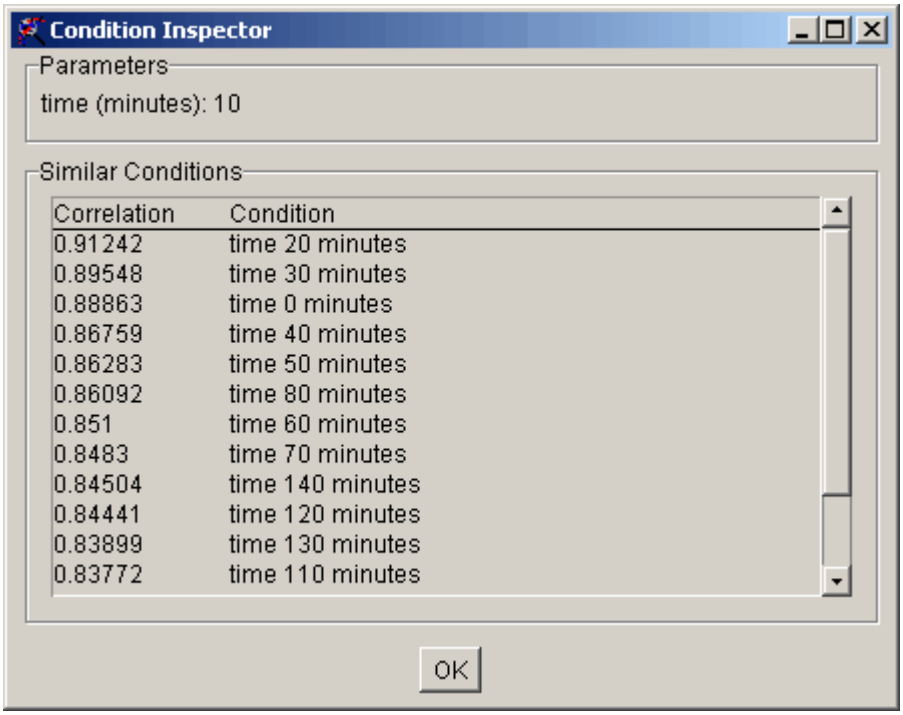


Figure 3-21 The Condition Inspector window

The Parameters Box

In this box is a brief description of the sample currently under inspection.

The Similar Conditions Box

- **Correlation—**This list of numbers is how closely correlated the other conditions in the experiment are to the one under inspection. The conditions are listed in the order from most closely correlating to least correlating.
- **Condition—**This is a list of the other conditions in this experiment, briefly described. Double-clicking any one of them will bring up a Condition Inspector for that condition.

List Inspector

You can view the contents of a gene list and the method with which it was created using the Gene List Inspector. The Gene List Inspector is especially useful in learning about lists that have been identified using the Similar List function. The Gene List Inspector shows the history of your gene list, a graph of your list, a table of all the genes included in the list, and a collection of gene lists that are statistically similar to your gene list.

The history of the gene list is in the upper left corner of the window. You can change this information with the **`Edit`** button. In the upper right corner of the window is a browser graphing your list. Right-clicking on this browser gives you several options (see “Using Genome Browser” on page 3-1 for information on browser options). In the center of the Gene List Inspector window is a table of all the genes included in the list. Double-clicking a gene or cell in this table brings up a Gene Inspector window for the selected gene (see “Gene Inspector” on page 3-37 for information

Copyright 1998-2001 Silicon Genetics

on the Gene Inspector). The Similar Lists box in the lower left of the window contains names of lists resembling the displayed list, or containing a statistically significant number of overlapping genes. The statistical significance is listed as the p-value for each of the similar lists. You can right-click on one of these lists to print or copy. Double-clicking a list brings up a Gene List Inspector for that list.

Figure 3-22 shows the Gene List Inspector window for the “like YMR199W (CLN1) (0.95)” list.

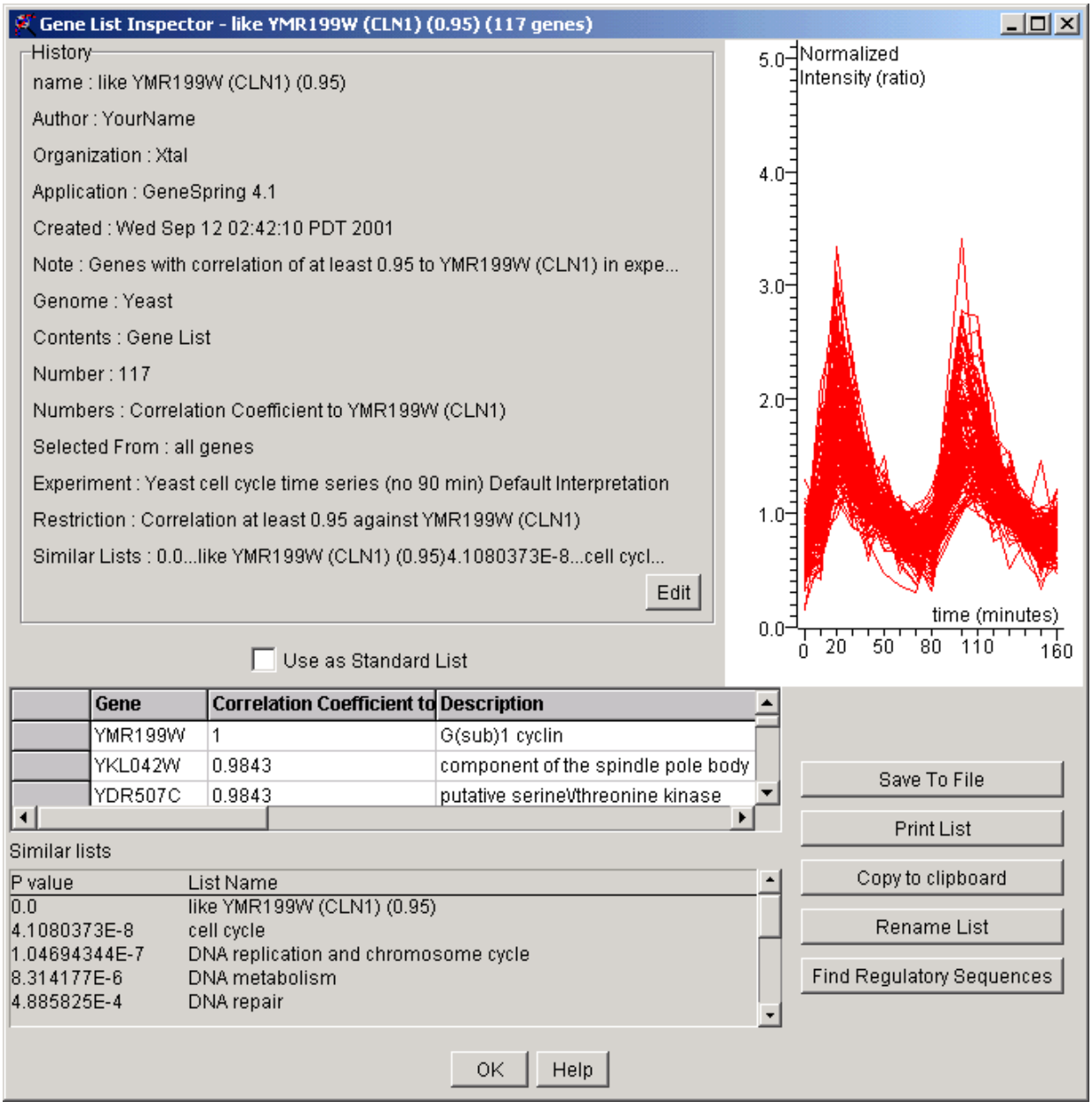


Figure 3-22 The List Inspector window

To use the Gene List Inspector

Double-click a gene list name in the navigator.

Or,

1. Right-click any gene list.
2. Select **Properties** from the pop-up menu.

To save the Gene List Inspector to a separate file, click the **Save to File** button. Select a directory and file name and click **Save**.

To print your list, click the **Print List** button. Click **OK**.

To copy a list, click the **Copy to clipboard** button. Paste into a text editor.

To rename a list, click the **Rename List** button. Type in the new name and click **OK**. You must also click **OK** in the main Gene List Inspector window to confirm your new name.

To use the Find Regulatory Sequences Function, see “Regulatory Sequences” on page 4-26 for information about the Find Potential Regulatory Sequences Window.

Classification Inspector

The Classification Inspector allows you to learn about the method used to construct a classification or to learn more about the variability explained by each class within a classification. To use the Classification Inspector, right-click a classification in the navigator panel and select the **Inspect** option.

Using the Classification Inspector Window

In Figure 3-23 the notes field contains information about the method used to make the classification. If the classification is the result of clustering, this field displays information such as the type of clustering, the distance metric, and the number of iterations that were used to perform the clustering. You can save your own comments about the classification here for future reference. The bottom half of the Classification Inspector contains a table with three columns:

- **Class:** the name given to each class
- **Genes:** the number of genes in each class
- **Average Radius:** the root mean square of the Euclidean distances between each gene and the centroid of each class. Classes with large radii are spread out and classes with small radii are tightly grouped.

Copyright 1998-2001 Silicon Genetics

At the bottom of the Classification Inspector window, the Percent Explained variability is displayed. This number is a measure of the quality of the classification; classifications in which the average radius of each class is small and in which the centroids of each class are located far apart from one another explain a high percentage of the total variability. GeneSpring expresses the percent explained variability, E, as:

$E = 100[G/(1+G)]$

Where G is the Calinski and Harabasz index of quality:

$G = [B / (c-1)]/ [W/ (n-c)]$

B is the sum of the squares of the distances between the cluster centroids and the mean of all genes in all classes, W is the sum of the squares of the distance between all genes and the centroid of the class to which the gene belongs, n is the total number of genes and c is the total number of classes.

This number is useful for comparing the quality of classifications that contain a different number of classes. The index of quality, G, takes into account the number of classes, so the quality will not rise limitlessly as the number of classes is increased. For example, a clustering method that produces six classes may explain 60% of the variability and one that produces 10 classes may explain 87% of the variability. However, when the number of classes is increased to 20, the percent of explained variability may drop, suggesting that 10 classes is a more effective classification than 20. Thus, the percentage of explained variability is useful in determining the optimum number of groups for a given clustering analysis.

Copyright 1998-2001 Silicon Genetics

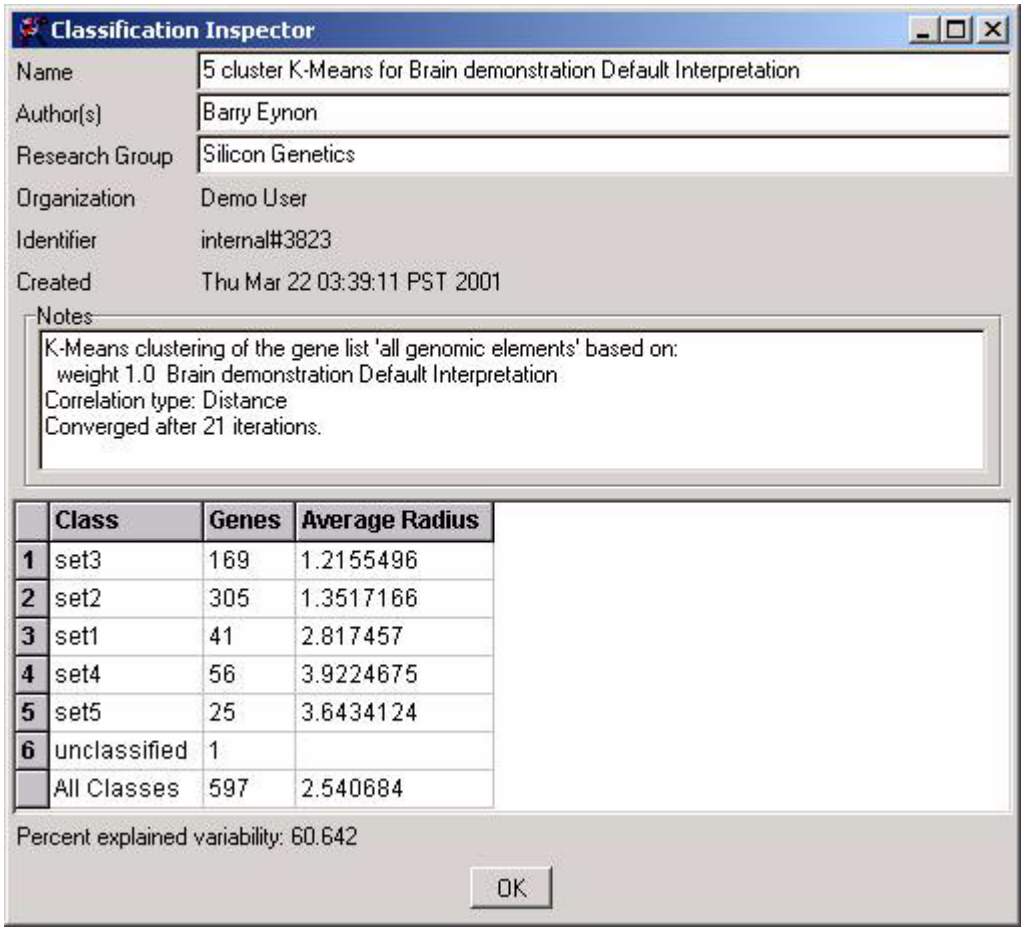


Figure 3-23 Classification Inspector for a k-means clustering with 5 groups

References for the Classification Inspector

Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics, 3, 1-27.

Gordon, A. D. *Classification*, 2nd Ed. Monographs on Statistics and Applied Probability 82. Chapman & Hall/CRC, Boca Raton (1999).

Copyright 1998-2001 Silicon Genetics

Chapter 4 Analyzing Data in GeneSpring

Filter Genes Analysis Tools

The Filter Genes Analysis tools allow you to take a current gene list and apply a series of restrictions (or filters) to make a smaller list. These restrictions can pertain to an entire experiment or interpretation, or to a single condition or sample. The filters include factors such as quality control, control strength, expression level constraints, sample to sample fold comparison, statistical group comparisons, and associated numbers restrictions. All restrictions applied to create a new list are saved as an attachment to the new list.

The ability to restrict a gene list based on the behavior of its genes in experiments or in individual samples is an important quality control tool. You may want to remove genes with low precision, large error values, those that do not vary significantly across multiple samples, or those with expression levels that are too close to the background. Filtering genes also allows you to search for genes that are differentially expressed over two or more conditions.

Filtering Genes

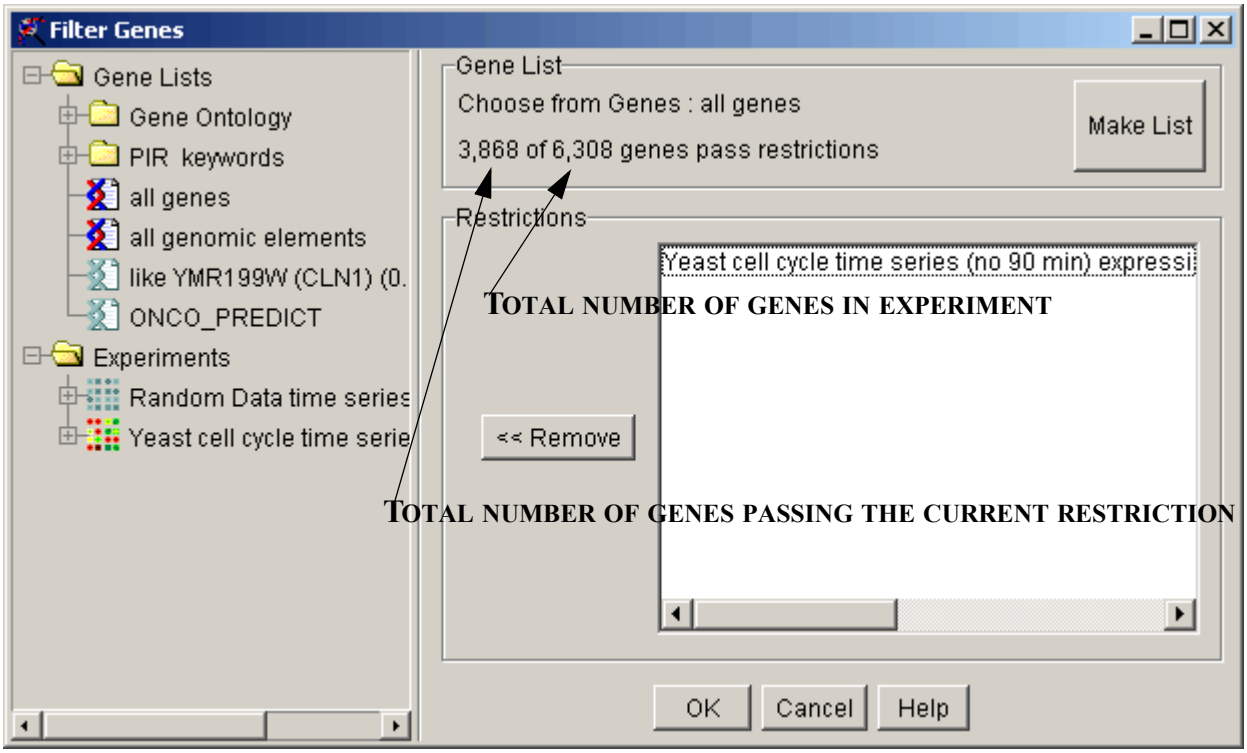


Figure 4-1 The Filter Genes window

1. Select **`Tools > Filter Genes`**. If you want to change the gene list, select a different gene list from the navigator panel of the Filter Genes window.
2. Right-click an experiment, sample or condition in the navigator.

Copyright 1998-2001 Silicon Genetics

3. Select one of the five restriction options available from the pop-up menu. You will be prompted for information about the type of restriction you want to make. There are five types of restrictions available:

   - **Expression Percentage Restriction** can be applied to entire experiments.
   - **Statistical Group Comparison Restriction** can be applied to entire experiments.
   - **Expression Restriction** can be applied to single conditions or samples.
   - **Condition to Condition Experiment Restriction** can be applied to single conditions or samples.
   - **Data File Restriction** can be used for either entire experiments or single conditions or samples.

   Details about the types of restrictions you can make are described below. A sixth option, **Inspect**, brings up the appropriate Inspector information window.

4. You can repeat steps 2 and 3 applying several restrictions at one time. To remove a restriction, click the text of the restriction in the Restrictions box and click the **Remove** button.

5. Click **OK** to make the list. Alternatively, click the **Make List** button to name and save the new list without closing the Filter Genes window, if you wish to continue applying filters.

6. Choose a name and destination folder for your new list and click **Save**.

Copyright 1998-2001 Silicon Genetics

Restrictions Over an Entire Experiment or Interpretation

Restricting by Expression Percentage

This restriction finds genes with certain values present in some of the conditions or samples in an experiment or interpretation. You can set what proportion of conditions must meet a certain threshold. For example, if you want to eliminate genes that do not meet a specified control value at least once in the experiment, you can filter them out by setting a minimum expression value to be met in at least one condition.

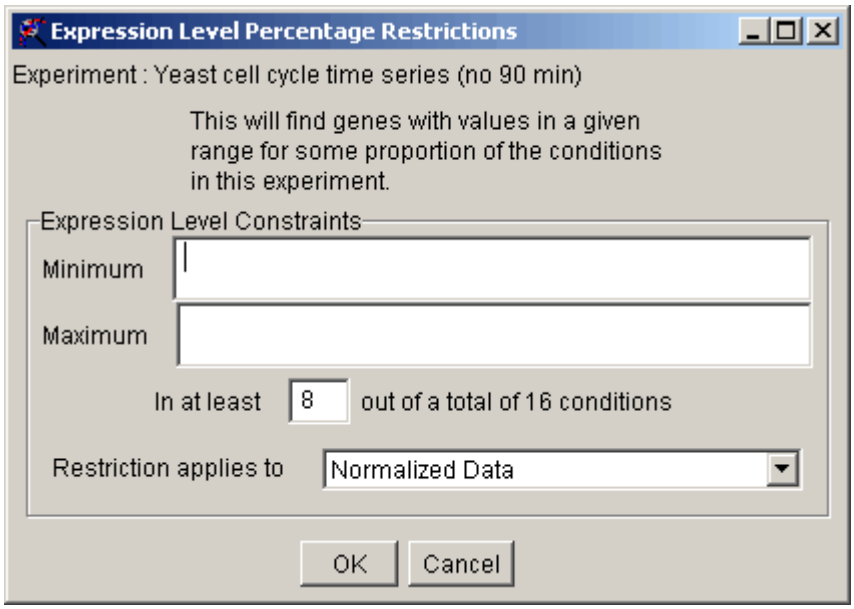

Expression Level Percentage Restrictions
Experiment : Yeast cell cycle time series (no 90 min)
This will find genes with values in a given range for some proportion of the conditions in this experiment.
Expression Level Constraints
Minimum
Maximum
In at least
8
out of a total of 16 conditions
Restriction applies to
Normalized Data
OK
Cancel

Figure 4-2 The Expression Level Percentage Restriction window

To perform an Expression Level Percentage Restriction, complete the following fields:

- **Minimum:** the smallest value any gene can have and GeneSpring will still allow it in your list (also known as the cut-off value).
- **Maximum:** the largest value any gene can have and GeneSpring will still allow it in your list.
- **In at least [ ] out of a total:** the number of conditions in the total experiment where genes must meet the specified requirements. This line can refer to the whole experiment. Adding any number where will cause GeneSpring to search every sample to determine if the gene passes.
- **Restriction applies to:** the type of data on which your restriction will be based. Please refer to “Data Types for Restrictions” on page 4-7.

Restricting by Statistical Group Comparison

The Statistical Group Comparison restriction finds genes with statistically significant differences in expression level between groups of samples. This restriction will remove genes based on the mean normalized expression levels of a group according to your current interpretation mode (logarithm, ratio or fold change). You will need to specify which parameter is to be used for the com-

 Copyright 1998-2001 Silicon Genetics

parison, the particular statistical test to be performed, and the cutoff on the p-value to be used in identifying statistically significant results.

For example, you can use the Statistical Group Comparison feature to filter out genes that do not vary significantly across different groups with multiple samples. This allows you to find those genes that exhibit important changes between various conditions of the experiment. This comparison is performed for each gene, and the genes with sufficiently small p-values are returned.

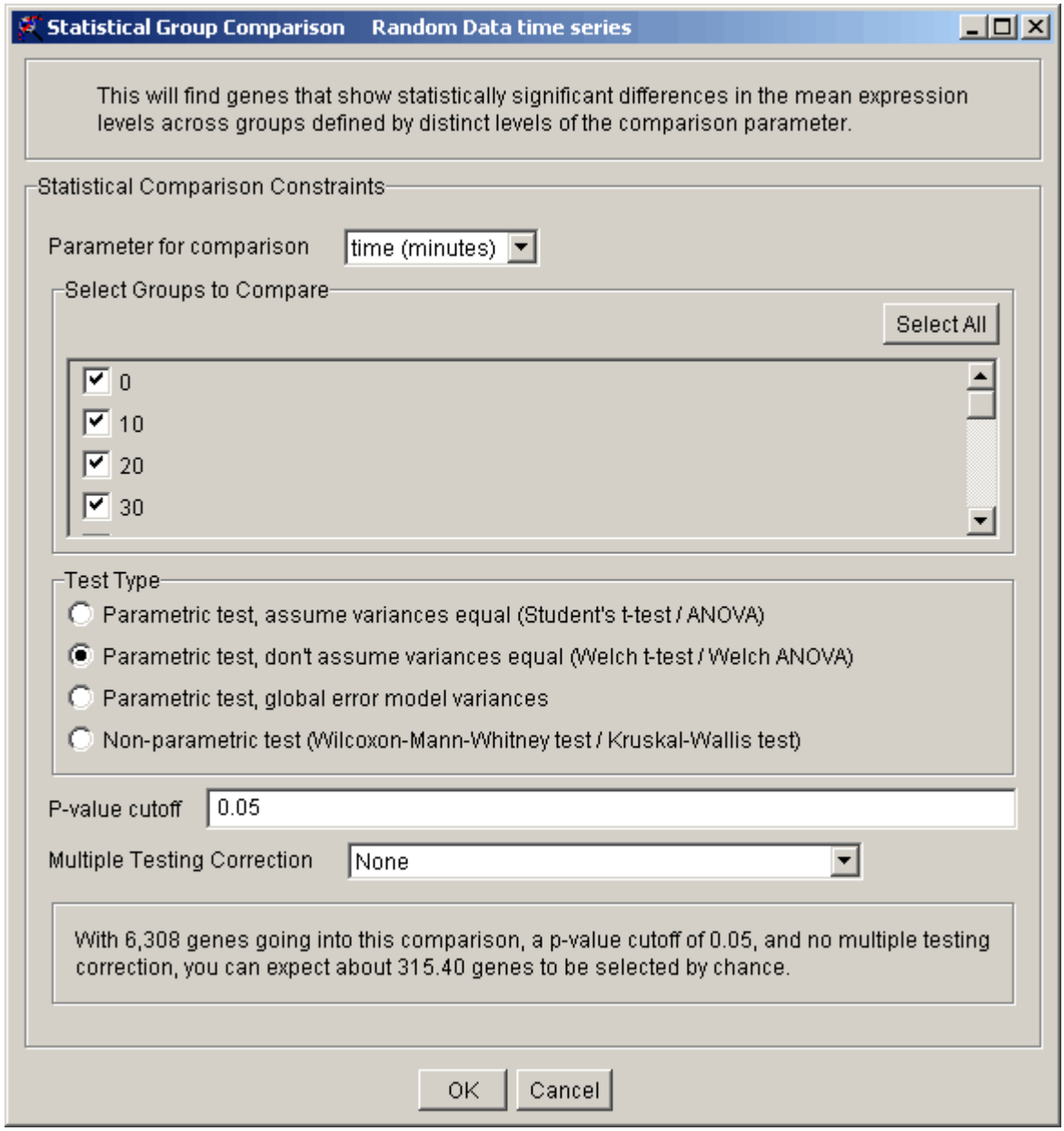

Statistical Group Comparison Random Data time series
This will find genes that show statistically significant differences in the mean expression levels across groups defined by distinct levels of the comparison parameter.
Statistical Comparison Constraints
Parameter for comparison
time (minutes)
Select Groups to Compare
Select All
0
10
20
30
Test Type
Parametric test, assume variances equal (Student's t-test / ANOVA)
Parametric test, don't assume variances equal (Welch t-test / Welch ANOVA)
Parametric test, global error model variances
Non-parametric test (Wilcoxon-Mann-Whitney test / Kruskal-Wallis test)
P-value cutoff
0.05
Multiple Testing Correction
None
With 6,308 genes going into this comparison, a p-value cutoff of 0.05, and no multiple testing correction, you can expect about 315.40 genes to be selected by chance.
OK
Cancel

Figure 4-3 The Statistical Group Comparison window

Copyright 1998-2001 Silicon Genetics

To Make a Statistical Group Comparison

1. Select the parameter on which you would like to base your comparison in the **Parameter for comparison** drop-down list.
2. Select the samples that you would like to compare by checking (or unchecking) the desired samples in the **Select Groups to Compare** box.
3. Select the type of test that you would like to perform. There are four testing options. For details on the formulae used for these tests see "Technical Details on the Statistical Group Comparison" on page N-1.
   - **Parametric test, assume variances equal** checkbox—filters based on the results of a Student's two-sample t-test for two groups or a one-way analysis of variance (ANOVA) for multiple groups.
   - **Parametric test, don't assume variances equal checkbox**—filters based on the results of an ANOVA or Welch's approximate t-test for two groups. This is the test that is most appropriate for standard experiments, when the global error model is not turned on or should not be used in the analysis.
   - **Parametric test, global error model variances**—filters based on the variances estimated by the global error model. If the global error model is not turned on, this test is equivalent to the Parametric test, don't assume variances are equal option.
   - **Non-Parametric test checkbox—**filters based on the rank of each sample, rather than the expression level. Non-parametric comparisons use the Wilcoxon two-sample rank test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups. This test will be most successful if you have more than five replicate samples in each group.
4. Select a minimum **P-value cutoff** for genes that pass the filter. Select a type of multiple testing correction. There are five options that are described below.

Multiple Testing Corrections

When testing the statistical significance of group comparisons for many genes, if you rely on the nominal p-value, many genes will pass the filter by chance alone. For instance, if you test 10,000 genes for reliable changes between groups at significance level 0.05, then (assuming the tests are independent) you would expect to misidentify about 500 genes as significant, even when there is no real difference gene expression. Even if you identify 1,000 genes showing significant behavior by this approach, *half* of the genes on the list will have appeared by chance, which lessens the value of the list. Multiple testing corrections adjust the individual p-value to account for this effect.

Suppose the p-value cutoff is α and the number of genes being tested is N. The first three procedures (Bonferroni, Holm, and Westfall and Young) control the family-wise error rate (FWER) which is the overall probability of obtaining even a single false positive test to be no more than α. This is a very strong criterion, but may be so strong for large lists of genes that no genes are identified as significant. The Benjamini and Hochberg test controls the false discovery rate, defined as the proportion of genes expected to be identified by chance relative to the total number of genes called significant.

Copyright 1998-2001 Silicon Genetics

- **Bonferroni:** The Bonferroni multiple testing correction, based on Bonferroni's inequality, limits the chance of a false positive results to be no more than α by multiplying each nominal p-value by *N* (with a maximum of 1). This process controls the FWER, and the expected number of genes by chance is α.
- **Bonferroni step-down (Holm):** The Holm step-down adjustment computes the most significant p-value, and whether it meets the α cutoff after multiplying by *N*. If that gene is found to be significant, then the next-most significant gene is considered, but the gene that was found significant is removed from the multiple-testing, so the multiple-testing adjustment is now based on N-1. This process is continued as long as genes pass the successive tests. This process controls the FWER, and expected number of genes by chance is α.
- **Westfall and Young permutation:** This procedure estimates the significance levels of each test by a nonparametric permutation calculation based on the distribution of the significance levels across all possible reassignments of samples to groups. For small numbers of permutations, all permutations are examined. If there are more than 1000 possible permutations, 1000 of them are selected randomly. P-values are evaluated with respect to this distribution using a step-down procedure as in the Holm procedure. This procedure controls the FWER, and the expected number of genes by chance is α. This test accounts for the dependence structure between genes, and should give a more powerful test than the Bonferroni or Holm procedure. However, the permutation process takes much longer to calculate.
- **Benjamini and Hochberg false discovery rate:** In contrast to the above procedures, the Benjamini and Hochberg procedure controls the false discovery rate (FDR), defined as the proportion of genes expected to occur by chance (assuming genes are independent) relative to the proportion of identified genes. Expected number of genes by chance is α times the number of tests found significant *after applying this correction*. There is no way to calculate this in advance, so the statement about the number expected will simply say expected number of genes by chance is 100α% of the genes identified. This procedure provides a good balance between discovery of significant genes and protection against false positives, since occurrence of the latter is held to a small proportion of the list, and will probably be the best choice of multiple-testing correction for most situations.

Copyright 1998-2001 Silicon Genetics

Restrictions over a Single Condition or Sample

Expression Restriction

The Expression Restriction finds genes with expression values that fall between specified minimum and maximum values for a particular condition. This tool is useful if you want to find genes that respond similarly to a given condition. For example, you may want to find genes in an inhibitor-treated sample with a minimum normalized expression of 3.

For details on the types of data you can apply this restriction to, please refer to “Data Types for Restrictions” on page 4-7.

Condition to Condition Comparison Restriction

The Condition to Condition Comparison Restriction finds genes based on a comparison between two samples or conditions. This tool is used to find fold changes in gene expression levels between two samples or conditions.

1. Select an individual sample or condition.
2. Right-click the sample or condition and select **`Add Condition to Condition Comparison Restriction`** from the pop-up menu. A window will open.
3. Open a second sample or condition from the Experiment menu in the mini-navigator of this window. Note that you have already selected the first condition to be compared.
4. From the pull-down menu choose whether you want the signal in the first sample or condition to be greater than, less than or equal to that in the second sample.
5. Enter a fold factor in the *by at least a factor of* field.
6. Select a type of data from the pull-down menu.

Data Types for Restrictions

You can change the type of data on which to base the restriction, by choosing from a drop-down list in the applicable window. Depending on which feature you are currently using, you may have access to only some of the options in the following list.

- **Normalized Data:** the values that GeneSpring displays in the Normalized column in the Gene Inspector.
- **Raw Data:** unnormalized experimental data. Note: if your computer is set for a default language that is not English, please make sure a consistent convention for decimal markers is followed.
- **Control Signal:** the normalization denominator.
- **Number of Replicates:** the number of samples in each condition.

Copyright 1998-2001 Silicon Genetics

- **Range of Normalized Data:** the difference between the minimum and maximum of the normalized data. You can use the Range of normalized data feature if you want genes with, for example, a compact range of data. This range refers to the variability in a single condition, not in the mean expression level over an entire experiment.

  NOTE: If your original data did not include measurement flags, you can use the **`Range of normalized data`** feature to filter out "Absent" genes by specifying a value 0 or above because Absent genes are not assigned any value.

- **Standard Error of Normalized Data:** the precision in an experimental condition as expressed in terms of standard error.

- **Standard Deviation of Normalized Data:** the precision in an experimental condition as expressed in terms of standard deviation. Silicon Genetics recommends three methods for filtering for reliable genes using the Standard Deviation of Normalized Data option:

  - If you want genes where the standard deviation of the individual normalized measure values is less than or equal to a maximum value, L, specify L as the maximum value.
  - If you want genes where the mean of the normalized values in each group has a standard error of L or less, specify L* sqrt(N) as the maximum value, where N is the number of replicates in each group.
  - If you want genes where the mean of the normalized values in each group is accurate to within +/-L with 95% confidence, then specify L *sqrt(N)/1.96 as the maximum value, again where N is the number of replicates in each group.

- **T-test probability:** the likelihood that the difference between the normalized expression level and normality (usually 1) is actually less than indicated.

Normalized, Control and Raw data are also displayed in the upper right corner of the Gene Inspector window.

Data File Restriction

The Data File Restriction allows you to filter genes based on values in a specific column of your experiment data files. For example, if you specified a flag column when you loaded your data, you can filter on Present or Marginal calls.

You can select any column name from your experiment from the Column drop-down menu. Alternatively, you can enter the column number in the Number box. If you have access to the original data files entered in GeneSpring, you can check them for column numbers. You can restrict the column values by choosing "greater than", "equal to" or "less than" from the pull-down menu and inserting a restriction value in the field provided.

For example, if you had loaded an Affymetrix file as your experiment, you could use the drop-down menu to select the Abs/call column and select for all entries equal to "M" if you wanted to make a list of just the marginal data.

Copyright 1998-2001 Silicon Genetics

Restricting by Associated Numbers

New in version 4.1 is the ability to restrict genes according to the numbers associated with them in a gene list. When you make a new list based on a filter or similarity metric, the value used as a filter will be associated with the genes on the new list. Some examples of associated numbers are correlation coefficients, p-values, fold change ratios, or in the case of a regulatory sequence search, the number of base pairs before the promoter region. Associated numbers can be found by double-clicking a gene list to bring up the Gene List Inspector.

Restricting genes by their associated numbers is useful if you want to use this information to create a more specific list of genes. For example, you may want to find genes that are highly similar to another gene (with a high correlation coefficient), or genes that are a specific distance from a promoter found using the Find Potential Regulatory Sequences tool.

Adding an Associated Number Restriction

1. Right-click the list with associated numbers in the Filter Genes window navigator. (This can also be accessed in complex correlations or clustering.)
2. Select **`Add Associated Numbers Restriction`**. You will see a new Associated Numbers Restrictions window.
3. Enter minimum and maximum restriction values in the fields provided and click **`OK`**.

The option is disabled if you right-click a gene list with no associated numbers. For example, this restriction cannot be applied to the “all genes” or “all genomic elements” lists because there are no associated values.

Changing a Restriction

When you double-click a restriction, GeneSpring will bring up a dialog box with the current restriction information. From there you can change any of the restrictions you defined. To apply the restriction to another experiment or another condition, you must begin again by right-clicking over that data-object in the mini-navigator and selecting a new restriction.

Once your list is made, GeneSpring will attach numbers to each gene in that list. These numbers can be seen using the Ordered List view or the List Inspector. Note that you can filter on any of these numbers. See “Adding an Associated Number Restriction” on page 4-9 for details on associated numbers.

Copyright 1998-2001 Silicon Genetics

References

Benjamini, Y. and Hochberg, Y. (1995) “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society B*, 57, 289 -300.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Department of Statistics Technical Report #578, University of California, Berkeley (http://stat-ftp.berkeley.edu/tech-reports/index.html)

Holm, S. (1979) “A Simple Sequentially Rejective Bonferroni Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65 -70.

Miller, R.G. (1981) *Simultaneous Statistical Inference*, Second Edition. New York: Springer-Verlag.

Westfall, P.H. and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*. New York: John Wiley & Sons, Inc.

Copyright 1998-2001 Silicon Genetics

New Gene List window

The *New Gene List* window is created when a new list is made. It allows you to accept or reject the list after seeing the genes it contains, and it allows you to set the name of the list. The example in Figure 4-1 is the result of doing a correlation to find all genes with a similar expression profile to YMR199W (CLN1).

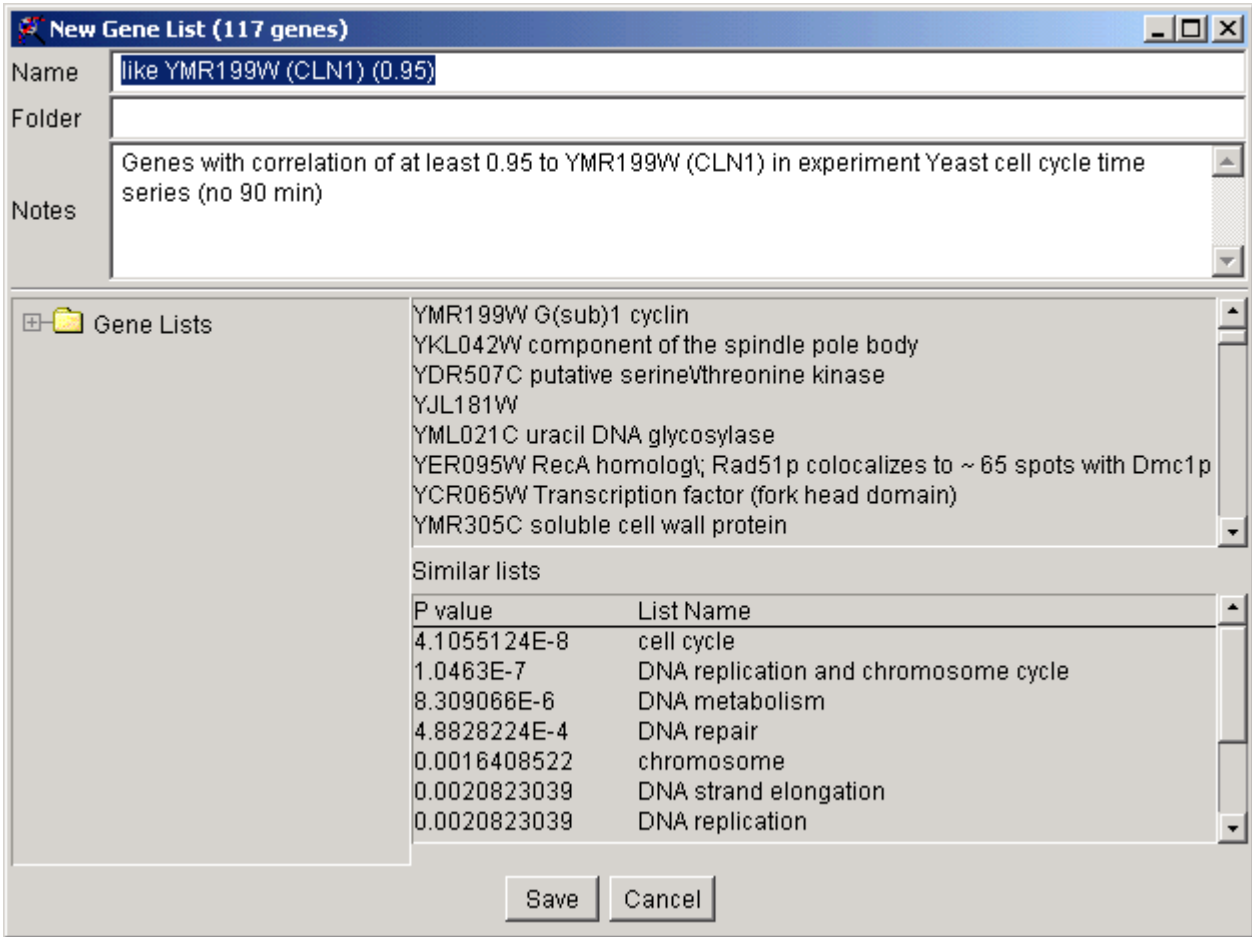


Figure 4-1 The "New Gene List" window

This list was a result of searching for all of the genes in the Yeast cell time series (no 90 min) experiment, having expression profiles within a .95 correlation to YMR199W (CLN1)'s profile. The genes fitting the restrictions of the search are listed in the top box. The lower box, titled *Similar lists*, contains the lists GeneSpring is aware of that are statistically similar to your new list. Similar means the lists contain a statistically significant overlapping of genes. How statistically significant the similarities are is given in the left-hand column of the bottom box, which lists the P-value (the probability of a false positive) for each of the lists in the right-hand column. (The p-value of a statistically significant list is at least 0.05.)

By double-clicking any item in the gene list or in the lists list, you will bring up an Inspector for the selected item.

Copyright 1998-2001 Silicon Genetics

Commands in the New Gene List window

- **Name:** The current (default) name is highlighted when the New Gene List window first appears, ready to be changed.
- **Save/Cancel:** Clicking the **Save** button saves the list and the name in the name box (in the example this name is "likeYMR19W (CLN1) (0.95)"), and also displays this list in the genome browser display. The **Cancel** button discards the list.
- **Inspecting a Gene in the Gene List Box:** Double-clicking a gene in the right-hand box brings up a Gene Inspector window, for that gene. See "Gene Inspector" on page 3-37 for a complete description of this window. The Gene Inspector window allows you to search the associated databases, to obtain more detailed information regarding a particular gene in the list.
- **Inspecting a List in the Similar Lists Box:** Double-clicking a list in the bottom box brings up a Gene List window displaying the genes in the selected list. This window is discussed in detail in "List Inspector" on page 3-44. The **OK** button and the **Cancel** button at the bottom of the Inspect Gene List window both exit the Inspect Gene List window, but do not close the New Gene List window.

Copyright 1998-2001 Silicon Genetics

Making Lists with the Find Similar Command

The Find Similar command allows you to do simple correlations, that is, to find genes with similar expression profiles to the gene currently being displayed. Similar genes have graphs with similar shapes.

Each gene expression profile must have the set minimum correlation to be considered similar. The higher you set the minimum correlation (maximum 1), the closer the gene expression profiles have to be.

To Make Lists with the Find Similar command

Double-click on a gene (this may be easier after zooming in)

Or,

1. Select **Edit** > **Find Gene**.
2. Enter in the name of your gene.
3. Press **Ctrl+I**.

This will take you to the Gene Inspector.

Then,

1. Specify the minimum correlation in the bottom left corner of the Gene Inspector window. Do this by placing your cursor in the box, highlighting the existing value, and then typing in your preferred value.
2. Click the **Find Similar** button. The New Gene List window will appear, which includes the genes in that list, as well as lists that are similar to your new gene list.

In views where lists can be ordered, such as the Ordered List view and Compare Genes to Genes view, lists made with the Find Similar command are ordered according to correlation coefficient, in descending order.

Copyright 1998-2001 Silicon Genetics

Making Lists with the Complex Correlation Command

The Complex Correlation command in the Gene Inspector allows you to set up complex correlations against the inspected gene. These correlations may involve more than one experiment or condition or extra restrictions on experiments.

To Make Lists with the Complex Correlation Command

1. Access the Gene Inspector by double-clicking on a gene (this may be easier after zooming in)

   Or,

   a. Select **Edit > Find Gene**.

   b. Enter in the name of your gene.

   c. Press **Ctrl+I**.

2. Click the **Complex Correlations** button in the bottom left corner of the Gene Inspector window. This will open the Multi-Experiment Correlation window.

3. Choose a gene list from the Gene List folder in the navigator by right-clicking the list and selecting **Set Gene List**.

4. To add an experiment or condition to the Correlations box, select the experiment or condition in the Experiments folder in the navigator and select the **Add** button under Correlations. Adding a new experiment or condition will bring up the New Correlation window.
   On the right side of the window is a cumulative distribution graph of the genes' correlations. The horizontal axis shows the correlation from zero to 1, the vertical axis depicts the number of genes. The green lines are your specified maximum and minimum values. If you change these values the green lines will move accordingly.

   a. The **Phase Offset** (series variable) function in the upper left corner of this window specifies how far the expression profiles should be offset in time (or other continuous parameter) from the expression profile of the gene to be correlated against. This function is optional. You can change the selected parameter to be offset by selecting a variable from the drop-down box.

   b. You can also select a weight for your experiment or condition, which is a measure of the influence the experiment or condition has on the correlation distance. For example, an experiment with a weight of 2.0 will be twice as influential as one with a weight of 1.0. For this equation, please see "The Correlations box" on page 4-16.

   c. You can also weight each gene by signal strength, with the result that each gene will have a different weight. To do this, click in the box marked **Weight by Control Strength**.

   d. Click **OK**.

   To remove an experiment or condition, click on the experiment or condition and select **Remove**.

5. Specify boundaries (correlation coefficients) for what is considered similar in the **Maximum** and **Minimum** boxes.

Copyright 1998-2001 Silicon Genetics

6. Choose a correlation from the drop-down menu. For more information about correlations, see “The Correlations box” on page 4-16 and “Equations for Correlations and other Similarity Measures” on page L-1.
7. The Restrictions box at the bottom of the window specifies the restrictions the genes have to pass before they reach the correlation stage. To add restrictions to the selected list, right-click an experiment or gene list in the navigator and select a restriction. For information on restrictions and how to apply them, see “Filtering Genes” on page 4-1.
8. Select the **`Make List`** button to make a list and keep the Multi-Experiment Correlation window open or the **`OK`** button to make a list and close the window. A New Gene List window will appear. This window lists all the genes in your new list as well as similar lists with their associated p-value.
9. Name your gene list and click **`Save`**. The list will show up in the Gene Lists folder of the main navigator.

The Multi-Experiment Correlation Window

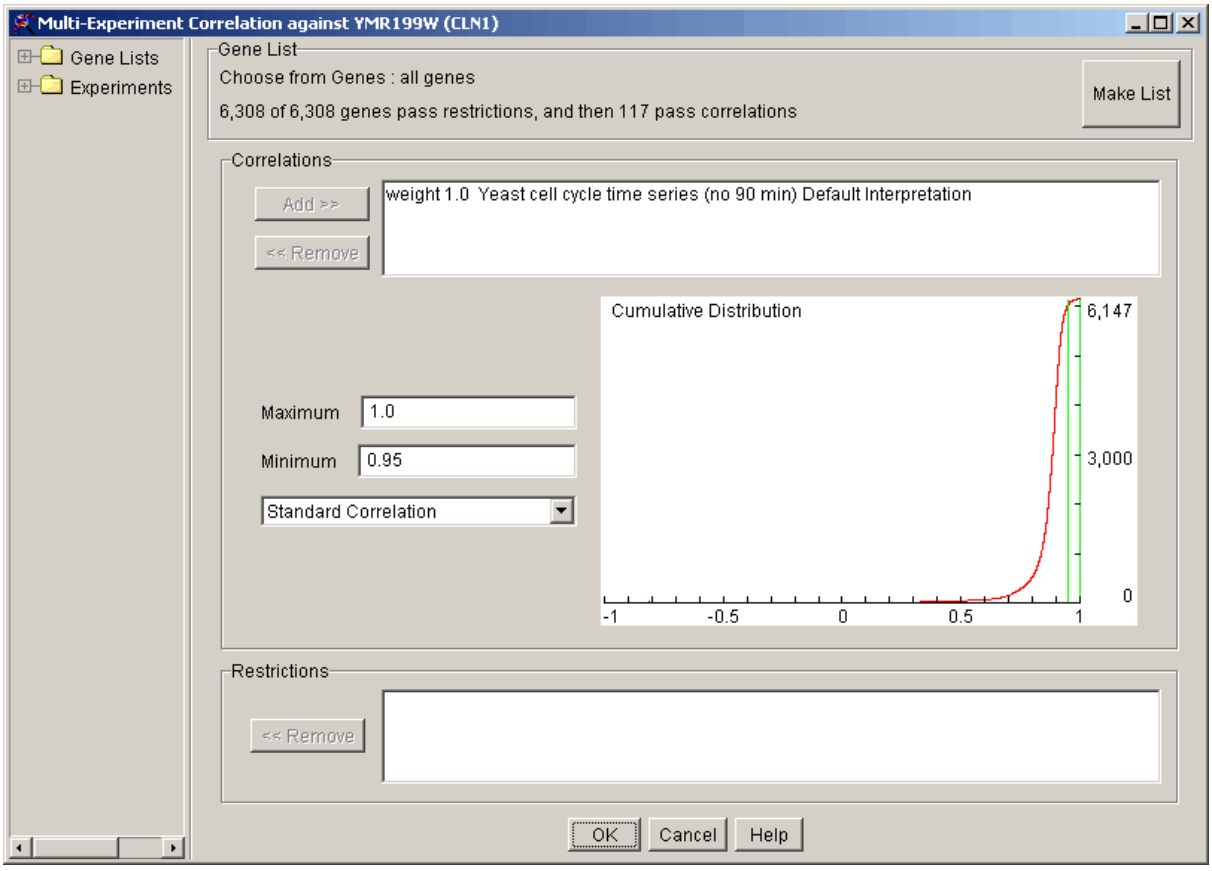


Figure 4-2 The Multi-Experiment Correlation window

Copyright 1998-2001 Silicon Genetics

The Correlations box

Below the Gene List box is the Correlations box.

On the left of the Correlations box is a white box indicating the experiments chosen to correlate against the gene listed in the title bar. The experiments selected may be weighted, making one more important than another. If both experiments chosen are given a weight of 1, they will be averaged equally. The name of the experiment is noted directly after its relative weight. The equation used to determine the overall correlation is:

$$X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}$$

- **A** is the correlation coefficient between the gene in question in experiment 1 and the gene named in the title-bar of the Multi-Experiment Correlation window, also from experiment 1.
- **a** is the weight specified for experiment 1.
- **B** is the correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- **b** is the weight associated with experiment 2.
- **C** is the correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- **c** is the weight associated with experiment 3.

and so on.

Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white Correlations box. If **X** is between the minimum and maximum correlations specified in the Multi-Experiment Correlation window, then the gene in question passes the correlations.

- **Standard Correlation:** Standard correlation measures the angular separation of expression vectors for Genes **A** and **B** around zero.
  Result = **a.b**/*(*|**a**||**b**|*)*
- **Smooth Correlation:** Make a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph screen. Do the same to make a vector **B** from **b**.
  Result = **A.B**/*(*|**A**||**B**|*)*
- **Change Correlation:** Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i)-\pi/4$.Do the same to make a vector **B** from **b**.
  Result = **A.B**/*(*|**A**||**B**|*)*
- **Upregulated Correlation:** Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\max(\text{atan}(a_{i+1}/a_i)-\pi/4,0)$. Do the same to make a vector **B** from **b**.
  Result = **A.B**/*(*|**A**||**B**|*)*

Copyright 1998-2001 Silicon Genetics

- **Pearson Correlation:** Calculate the mean of all elements in vector **a**. Then subtract that value from each element in **a**. Call the resulting vector **A**. Do the same for **b** to make a vector **B**.
  Result = **A.B**/*(*|**A**||**B**|*)*
- **Distance:** Distance is not a correlation at all, but a measurement of dissimilarity. Distance is the measurement of Euclidian distance between the expression profile for gene **A** (defined by its expression values for each point in N-dimensional space, where N is the number of conditions with data in your experiment) and the expression profile for gene **B**.
  Result = |**a-b**| divided by the square root of the number of conditions with data
- **Spearman Correlation:** Order all the elements of vector **a**. Use this order to assign a rank to each element of **a**. Make a new vector **a'** where the i^{th} element in **a'** is the rank of a_i in **a**. Now make a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, make a vector **B** from **b**.
  Result = **A.B**/*(*|**A**||**B**|*)*
- **Spearman Confidence:** Compute a value *r* of the spearman correlation as described above.
  Result =1-(probability you would get a value of *r* or higher by chance.)
- **Two sided Spearman Confidence:** Compute a value *r* of the spearman correlation as described above.
  Result =1-(probability you would get a value of |*r*| or higher, or -|r| or lower, by chance.)

The Restrictions box

The bottom white box is labeled *Restrictions*. In it are the restrictions the genes have to pass before they reach the correlation stage. The possible restrictions are discussed in detail in “Filtering Genes” on page 4-1.

Creating and Saving Your Correlated List

The *Make List* command makes a list but does not close the Multi-Experiment Correlation window. The **OK** button, at the bottom of the window, makes a list and closes the Multi-Experiment Correlation window. The **Cancel** button, also at the bottom of the window, simply closes the Multi-Experiment Correlation window.

Type in a unique name for your new list in the *Name* box and click **OK**.

Copyright 1998-2001 Silicon Genetics

Finding Offset Genes

In GeneSpring you can find genes whose profiles are similar to a specific gene, but are offset by one or more conditions.

1. Start from the Gene Inspector window. Zoom in on any gene using the **`Edit > Find Gene`** and double-click (or **`Ctrl+I`**).
2. Click the **Complex Correlations** button in the lower left corner of the window. For details about the other elements in the Gene Inspector window, please refer to “Gene Inspector” on page 3-37.
3. Double-click the **experiment name** in the *Correlations* box at the center of the window. This will bring up the *New Correlation* box with the default settings of that experiment. This is the same box you would see if you added a new experiment to this correlation. In the *phase offset* section, you will need to select a parameter from the drop-down list. You will also need to enter a number to offset from. What number you will enter depends on what makes sense with your chosen parameter.
4. Click **`OK`**. This will return you to the previous window, the Multi-Experiment Correlation window. Click the **Make List** button in the upper-right corner of the window.

GeneSpring will now look for genes with a similar shape to the inspected gene, but offset according to your input.

When GeneSpring has found genes whose profiles were similar but offset from your inspected genes, a New Gene List window will appear. Use the New Gene List window to name and save your new list. This feature can be used if you want to see what genes might have triggered activity.

Copyright 1998-2001 Silicon Genetics

Making Lists from Properties

You can make gene lists based on the properties (annotations) contained in your Master Gene Table. Such lists are not ordered.

To Make Lists from Properties

1. Select **`Annotations > Make Gene List from Properties`** (pre-4.1 users select **`Tools > Make Gene List from Properties`**).
2. Choose a property from the pull-down menu on which to base your list.
3. Deselect the Divide by semicolons checkbox if you do not want your data separated by semicolons.
4. You can tell GeneSpring to include a list only if it has a certain number of members, or you can include all lists. By default, GeneSpring removes gene lists with one or fewer members. Change this number in the text box provided, or include everything by deselecting the Remove classifications with 1 or fewer checkbox.
5. Under *Call Classification,* name your gene list folder.
6. Click **`OK`**. A new folder with the gene list you created will appear in your Gene Lists folder.

Making Lists with the Venn Diagram

A Venn Diagram allows you to quickly visualize genes common to more than one gene list. You can also find genes present in a specific list only. The gray area behind the circles represents the Venn Diagram "universe" (the selected gene list). Genes in the selected list that are common to gene lists represented by the Venn diagram circles appear as numbers in those circles. For information about creating and filling Venn Diagrams, see "Color by Venn Diagram" on page 3-33.

To Make a list with the Venn Diagram

1. Right-click the area of the Venn Diagram in which you would like to make a list.

   Select an option from the pop-up menu. A New Gene List window will appear. If you click in an area where two circles overlap, you will have the following options:

   - **Make list of these genes:** lists genes in the immediate geometric area.
   - **Make list of genes in both lists:** lists genes common to the two circles, i.e. the intersection.
   - **Make list of genes in either list:** lists all genes in the two circles, i.e. the union.

   If you click in an area where three circles overlap, you will have the following options:

   - **Make list of genes in all lists:** lists genes common to the three circles, i.e. the intersection.
   - **Make list of genes in any list:** lists all genes in the three circles, i.e. the union.

Copyright 1998-2001 Silicon Genetics

If you click a non-overlapping (gray) area, you can make a list of genes in that section only.

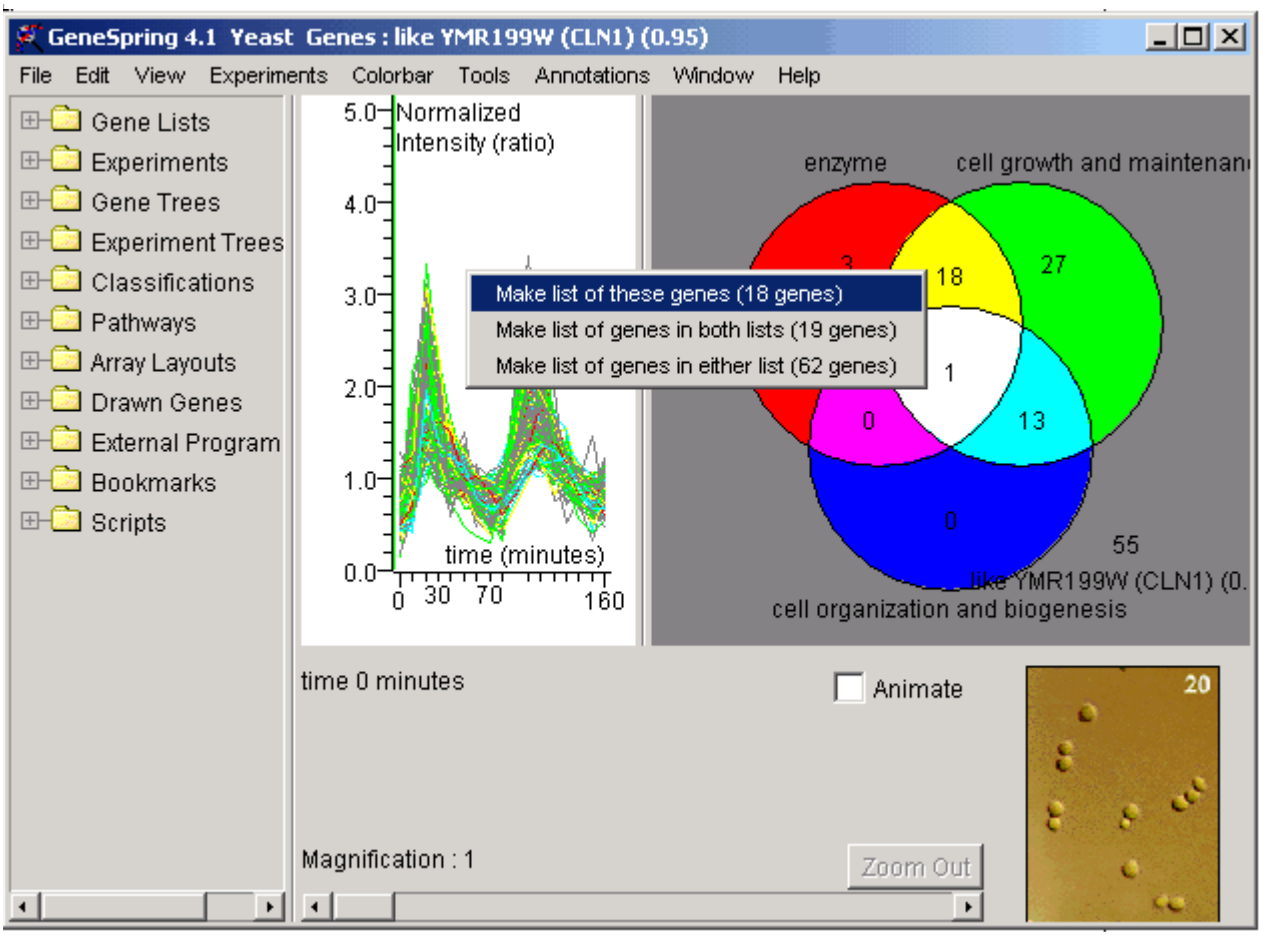


Figure 4-3 A Venn diagram with pop-up menu

2. Name and save your new list.

In views where lists can be ordered, such as the Ordered List view and Compare Genes to Genes view, lists made from the Venn diagram are ordered according to the values associated with the lists you used to create the Venn Diagram. When more than one of these lists has values, genes are ordered according the values of the last list added to the Venn diagram when it was created.

Copyright 1998-2001 Silicon Genetics

Making Lists from Classifications

You can generate gene lists from any classification. For example, if you have a 5-cluster k-means classification, you can view which genes are in each cluster by making a gene list from the k-means classification.

To make a Gene List from a Classification

1. Right-click a classification in the Classifications folder in the navigator.
2. Select **`Make gene lists`**. GeneSpring will create a gene list folder for the classification containing one list for each cluster. You will find this folder in the Gene Lists folder in the navigator.

Find Interesting Genes

The Find Interesting Genes command finds genes that have gone through the largest expression changes during the experiment and have high trust values.

To find Interesting Genes

1. Select **`Tools > Find Interesting Genes`**. A dialog box will appear showing one of the most interesting genes in your experiment.
2. Click the button in the box. The Gene Inspector for that gene will appear. (See “Gene Inspector” on page 3-37 for information about the Gene Inspector.) To find more interesting genes, repeat these steps.

The Find Interesting Genes command also automatically creates a list of interesting genes, complete with an interest score for each one, in your Gene List folder.

In views where lists can be ordered, such as the Ordered List view and Compare Genes to Genes view, lists of interesting genes are ordered according to interest score, in descending order. For an example of Ordered List view, please refer to “Ordered List View” on page 3-21.

Copyright 1998-2001 Silicon Genetics

Making Lists from Selected Genes

This command allows you to make lists from genes you select graphically.

To make a list from selected genes

There are two ways to select a set of genes. If genes are grouped together in the browser, you can select a set in the same way you select an area to enlarge:

1. While holding down the shift key, click and drag a rectangle across the region you wish to select.
2. Release the cursor while continuing to hold down the shift key. Selected genes will appear in white.

   Or,

- Select multiple genes by clicking over their representative lines or rectangles while holding down the shift key.

Once you have selected all the genes you want in your new list, right-click in the genome browser and select **`Make List from Selected Genes`** from the pop-up menu. A New Gene List window will appear. Name your list and click **`Save`**. For more information about this window, see "New Gene List window" on page 4-11.

Creating Drawn Genes

The Creating Drawn Genes function allows you to draw a pseudo-gene to represent a hypothetical expression pattern. This function is useful if you have some idea of what gene expression pattern you are looking for, as you can simply draw a pattern and look for genes that behave similarly. You must be in Graph, Bar Graph, Scatter Plot, or Graph by Genes view to create a drawn gene. Double-clicking on the drawn gene will open the Gene Inspector for that gene.

To create a drawn gene

1. Select **`Tools > Show Drawable Gene`**. A new gene will appear on the screen, at the normalized median of your data (usually 1.0).
2. To change the shape of this gene, click on the gene and drag it while holding down the control key.
   - **Mac Users:** Please use Option-Click to alter your Drawn Gene.

To save a Drawn Gene

1. Double-click the drawn gene to open the Gene Inspector.
2. Click the **`Save As Drawn Gene`** box in the bottom left of the window.
3. Give your new profile a name and click **`Save`**. Your new drawn gene will appear in the Drawn Genes folder in the navigator.

Copyright 1998-2001 Silicon Genetics

To make Lists from Drawn Genes

1. Double-click the drawn gene to open the Gene Inspector.
2. Click the **Find Similar** button in the bottom left corner of the window. A New Gene List window will appear with a list of similar genes and lists.
3. Name your list and click **Save**. Your new list will appear in the genome browser and in your Gene Lists folder.

Pathways

A pathway is a graphical representation of the interaction between gene products in a biological system. Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context. You can zoom in on a pathway, and move the slider to watch gene expression change over the experimental conditions.

You can draw pathways yourself or use publicly available pathways such as KEGG (Kyoto Encyclopedia of Genes and Genomes). One scenario in which a pathway can be very useful is if you are trying to identify a class of genes that are associated with a particular step or regulatory element within a pathway.

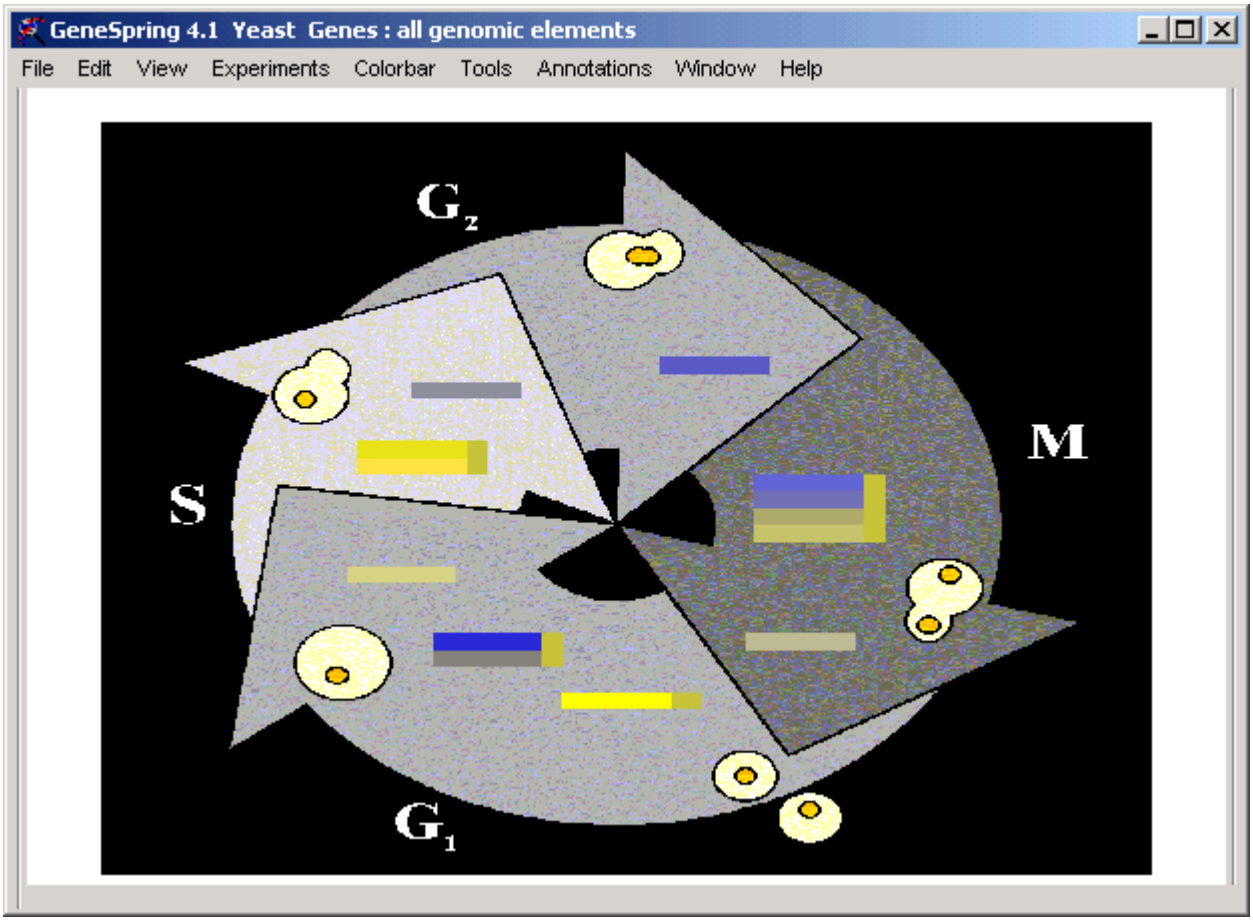


Figure 4-4 A Pathway cyclin and other genes during Metaphase of the cell cycle

Copyright 1998-2001 Silicon Genetics

In Figure 4-4, at about 20 minutes, you can see that the genes believed to be involved in S phase are overexpressed (colored in red).

Importing a Pathway

You can find pathways on the Web at sites such as:

- KEGG at ftp://kegg.genome.ad.jp/pathways/
- BioCarta at www.biocarta.com
- SPAD (Signaling Pathway Database) at http://www.grt.kyushu-u.ac.jp/spad/menu.html.

To import a pathway, your pathway image must be in a *.gif* or *.jpeg* file format. You can manually import the file into GeneSpring by placing it in Program Files/SiliconGenetics/GeneSpring/Data/YourGenome/Pathways), or by doing the following:

1. Select **`File > Open Genome or Array`** and choose the genome in which you want to place the pathway.
2. Select **`File > New Pathway`**. The Select Image File dialog box will appear.
3. Browse for your image file and select it. Click **`Open`**. This will bring up the Choose Pathway Name window.
4. Enter a name for your pathway and folder and click **`Save`**. You can now find your pathway in the Pathways folder in the navigator.

Adding a Gene to a Pathway

Once you have successfully imported your graphics file into GeneSpring, you are ready to place genes on top of the background image.

1. Open the appropriate Pathway in the navigator.
2. While holding down the **`Ctrl`** key, draw a box where you would like the gene to appear on the pathway. (Mac users should press Option and drag the mouse.) The New Genes on Pathway window will appear.
3. Type in the gene name, accession number, or keyword (such as a word in a gene's descriptor) and click **`OK`**. The gene name should now appear on the pathway. To enter multiple genes in one location, separate gene names or keywords with semicolons.
4. If the gene name or keyword is present for more than one gene, another window will appear directing you to choose a gene ID from a list. Double-click on the appropriate ID.

If you make a mistake, you can right-click on the gene you would like to remove and select **`Delete Pathway Element`**.

Copyright 1998-2001 Silicon Genetics

Adding KEGG Pathways

When you import a pathway from KEGG (Kyoto Encyclopedia of Genes and Genomes), GeneSpring can use the associated *.html* file to add relevant genes to the pathway. Because GeneSpring locates these genes by EC number, you need to have the EC numbers for your genes in your genome. You can automatically retrieve these numbers from GenBank and LocusLink using GeneSpider.

To obtain the necessary KEGG files:

1. Point your Internet Explorer or an FTP client to ftp://kegg.genome.ad.jp/pathways/.
2. Copy and paste the map folder (which contains organism-independent pathways) into the Pathways folder in the selected genome (e.g., Program Files/SiliconGenetics/GeneSpring/Data/YourGenome/Pathways).

The folders that correspond to organism-specific pathways are not always recognized by GeneSpring because the annotation for some genes is in a modified format.

Finding New Genes on a Pathway

GeneSpring uses proprietary algorithms to predict the genes that fit near a selected point on a pathway. After you select a point, GeneSpring makes two lists of genes from those currently displayed on your diagram. List A contains the two genes that appear closest to your selected point on the diagram and list B contains all other genes on the pathway.

GeneSpring then examines all the genes on your currently selected gene list and finds all genes whose minimum similarity (correlation) with genes on list A is higher than their maximum similarity with genes on list B. These genes are made into a separate list for you to examine. You can place a gene from this list on the pathway (see “Adding a Gene to a Pathway” on page 4-24).

Note that if your pathway geometry is complex, this procedure will not be particularly useful as it relies on screen distance only, not pathway structure or connectivity.

To Find New Genes on a Pathway

1. Right-click near a group of genes displayed on your pathway.
2. Choose the option **`Find Genes Which Could Fit Here`**. The New Gene List window will appear.
3. Enter a name and folder for your gene list and click **`Save`**. Your new gene list will be saved in your Gene Lists folder.

Pathway Commands

Right-click your Pathway in the navigator for the following options:

- **Display Pathway:** Displays the selected pathway in the genome browser.
- **Properties:** Brings up the Properties box listing such details as pathway history and genome.
- **Attachments:** Allows you to add a text or picture attachment to your Pathway

Copyright 1998-2001 Silicon Genetics

- **Make Gene List:** Allows you to save a list of all the genes on the selected pathway.
- **Publish to GeNet:** Uploads your information and the pathway picture to GeNet (see "Publish to GeNet" on page 6-6.
- **Delete Pathway:** Lets you delete a pathway. A confirmation dialog box appears.
- **Rename Pathway:** Allows you to rename your pathway

Regulatory Sequences

The Find Potential Regulatory Sequence window allows you to find common regulatory sequences within genes in a gene list or to search for a known sequence. It also compares the frequency of occurrence against all other gene lists in the genome.

This feature is useful for finding genes sharing similar regulatory sequences or having a particular regulatory sequence in common.

When the regulatory sequences tool compares genes to the remainder of the genome, it uses the "all genes" list. The "all genomic elements" list includes non-gene elements that are not expressed.

In GeneSpring version 4.0 and later, the sequence information will be loaded automatically. Note: You can change the load automatically feature by going to **`Edit > Preferences > Genome/Array View`** and remove the check from the Load Sequence checkbox.

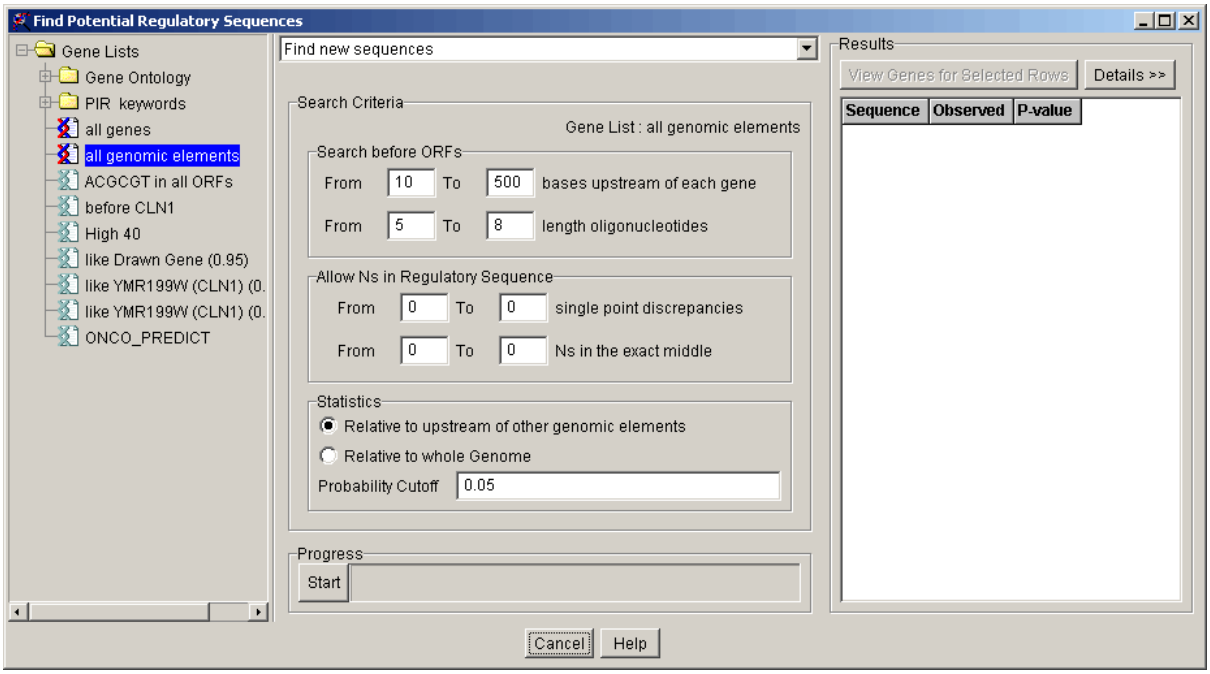


Figure 4-5 The Regulatory Sequences window

Copyright 1998-2001 Silicon Genetics

To find a Potential Regulatory Sequence

1. Select **`Tools > Find Potential Regulatory Sequences`**. The Find Potential Regulatory Sequences window will appear.
2. Select a gene list from the Gene Lists folder in the mini-navigator of the window. Note: Do not choose the "all genes" or "all genomic elements" gene lists because you are already comparing your selected gene list against all other genes in the genome.
3. Choose **`Find new regulatory sequence`** or **`Enter a specific regulatory sequence`** from the pull-down menu at the top center of the window.
   - **Find new regulatory sequence:** This option searches for short sequences upstream of the genes in the current gene list or across the entire genome.
   - **Enter a specific regulatory sequence:** This option allows you to enter a known sequence.
4. Enter the number of bases upstream of each gene you would like to search in the Search Before ORF section of the window. For example, if you enter "From 10 To 100" on a search for ACGCGT, GeneSpring will search for any part of the promoter within the region between 10 and 100. The smaller the range between these numbers, the more likely the results will be statistically significant. Larger sequences may take longer to search. You can also search for common sequences within the ORF by using negative numbers for the bases.
   - Enter the length of the oligonucleotides to search for if you have selected the **`Find new regulatory sequence`** option in the first step.
   - Enter the promoter sequence in the Enter Sequence textbox if you have selected **`Enter a specific regulatory sequence`** in the first step.
5. Enter the number of single point discrepancies allowed in the textbox provided. This refers to a maximum number of mismatches allowed, i.e., if you specify 1 single point discrepancy, then ACGCGAT satisfies a search for ACGCGTT.
6. Enter the range of base gaps in the exact middle if you have selected the **`Find new regulatory sequence`** option in the first step. This refers to the size of an allowable hole in the middle of the sequence, allowing you to look for sequences such as ACGnnnCGT, which is biologically relevant due to loops and non-binding areas. The gap must be in the exact middle, with the longer side of odd sequences appearing before the Ns. The gap does not count towards the sequence length specified; hence ACGnnnCGT would be returned as an oligonucleotide of length 6.
7. Select whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence. The first option is far more common.
   - The Probability Cutoff textbox indicates the level of significance (P-value) needed for an oligomer to be listed in the results. You may change this value if you wish.
8. Select the **`Search`** button. The button will change to a **`Stop Search`** button. The progress bar will lengthen as your search progresses.

Copyright 1998-2001 Silicon Genetics

Viewing Regulatory Sequence Search Results

The search results will be shown on the right-hand *Results* area of the Find Potential Regulatory Sequences window. Selecting the **`View Details`** button provides expanded results data that can be viewed by scrolling. Selecting the **`View Genes for Selected Row`** button brings up the *Conjectured Regulatory Sequence* window. Double-clicking any of the sequences in the table on the left brings up the Conjectured Regulatory Sequence window.

- **Sequence:** The nucleotide sequence of the oligomer.
- **Observed:** The number of genes in the list where the oligomer was found.
- **P-value:** The probability (P-Value) that the number of occurrences in the list came about by chance. Only nucleotide motifs with P-values below the specified probability cutoff (in this case 0.05 or 5%) are shown.
- **Random Rate:** The intrinsic probability, which is the percent of genes you would expect this specific nucleotide combination to appear upstream of, if the nucleotide sequence were strictly random (it is not, of course, but this is a good value to compare the observed probability to).
- **Observed—Other Genes:** The observed probability of this sequence motif appearing upstream of genes other than the list under inspection. If the option **`Relative to sequence upstream of other genes`** is selected, this becomes the probability of the observed sequence occurring relative to the genes not in the list, i.e., relative to the "all genes" list.
  If the option **`Relative to whole genomic sequence`** is selected, this becomes the probability of one or more occurrences of the sequence based on the rate of occurrence in the entire genome.
  The formula used to calculate this is:
  $1-(1-{}^{k}/_{b})^{n}$
  where k = the number of occurrences in the whole sequence
  b = the total number of bases
  n = the length of the upstream region being searched
- **Expected:** The number of incidences in the searched gene list, that you would expect this oligomer to occur. The number for the *Expected* column is derived using the larger of the intrinsic probability and the observed probability values.
- **Single P:** this column gives the Single P value for the motif. This is the chance this particular sequence would be found if only one test was performed.
- **Tests:** The number of tests run to come up with these motifs is given in the last column. This is the number of oligomers tested that were the length of the sequence motif found.

Copyright 1998-2001 Silicon Genetics

Using the Conjectured Regulatory Sequence window

The Conjectured Regulatory Sequence window displays the common nucleotide sequence, showing the 10 bases that precede and follow it in the area near (or in) each gene where the oligomer is found. It also gives a brief description of the statistics listed in the Results box of the Find Potential Regulatory Sequences window, and allows you to modify the observed motif by removing an item, extending the promoter or making a new gene list.

Double-clicking one of the sequence motifs given in the Results box of the Find Potential Regulatory Sequences window will bring up the Conjectured Regulatory Sequence window.

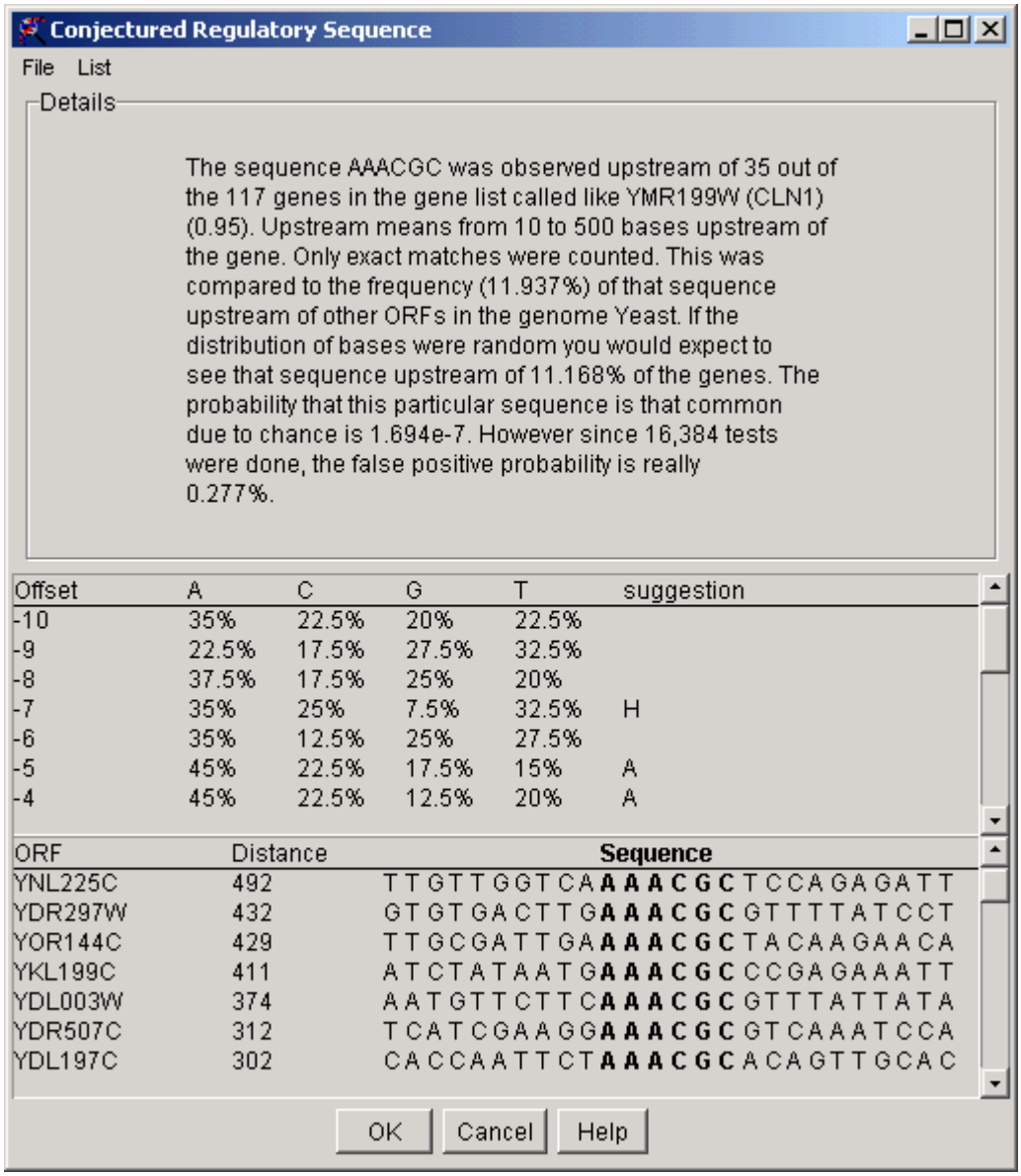

Conjectured Regulatory Sequence
File List
Details
The sequence AAACGC was observed upstream of 35 out of the 117 genes in the gene list called like YMR199W (CLN1) (0.95). Upstream means from 10 to 500 bases upstream of the gene. Only exact matches were counted. This was compared to the frequency (11.937%) of that sequence upstream of other ORFs in the genome Yeast. If the distribution of bases were random you would expect to see that sequence upstream of 11.168% of the genes. The probability that this particular sequence is that common due to chance is 1.694e-7. However since 16,384 tests were done, the false positive probability is really 0.277%.
Offset A C G T suggestion
-10 35% 22.5% 20% 22.5%
-9 22.5% 17.5% 27.5% 32.5%
-8 37.5% 17.5% 25% 20%
-7 35% 25% 7.5% 32.5% H
-6 35% 12.5% 25% 27.5%
-5 45% 22.5% 17.5% 15% A
-4 45% 22.5% 12.5% 20% A
ORF Distance Sequence
YNL225C 492 TTGTTGGTCAAAACGCTCCAGAGATT
YDR297W 432 GTGTGACTTGAAAACGCGTTTTATCCT
YOR144C 429 TTGCGATTGAAAACGCTACAAGAACA
YKL199C 411 ATCTATAATGAAAACGCCCGAGAAATT
YDL003W 374 AATGTTCTTCAAAACGCGTTTATTATA
YDR507C 312 TCATCGAAGGAAAACGCGTCAAATCCA
YDL197C 302 CACCAATTCTAAAACGCACAGTTGCAC
OK
Cancel
Help

Figure 4-6 The Conjectured Regulatory Sequence window

Two drop-down menus, **File** and **List** are located at the top of the window.

- **File:** Contains two commands: **Print** and **Close**.
  - **Print:** Prints the list in the lower half of the Conjectured Regulatory Sequence window.
  - **Close:** Closes the Conjectured Regulatory Sequence window.
- **List:** Contains three commands: **Remove Item**, **Make Gene List**, and **Extend Promoter**.
  - **Remove Item:** Removes the highlighted item and its associated sequence motif from the list matching the common sequence motif being examined.
  - **Make Gene List:** Brings up the new Gene List window for you to name and save a new gene list. When a gene list is produced based on the occurrence of a specified sequence (in this example, ACGCG in the yeast data) there is a number associated with each gene corresponding to distance of the first such sequence upstream of the ORF. The numbering begins from first nucleotide. These numbers can be easily viewed by zooming in on the Ordered list view or opening the Gene List Inspector.
  - **Extend Promoter:** Adds a new, longer and hopefully better promoter in the Find Potential Regulatory Sequences window.
- **Details box:** This box gives a general description of the common sequence motif being inspected. The details found in this box are the same numbers listed in the right-hand columns of the Results box in the Find Potential Regulatory Sequences window.
- **The Offset Bases box:** The middle third of the Conjectured Regulatory Sequence window contains statistics on the bases to either side of the motif. The first column gives the offset from the observed sequence. The next four columns give the percentage of genes with that base in that position. The last column contains a suggested extension to the motif.
- **ORF Box:** The bottom third of the Conjectured Regulatory Sequence window contains the sequence information for the motif being inspected, as it occurs in the nucleotide sequence in the area near (or in) each gene where it is found. There are three columns of data.
  - **ORF:** This indicates the gene that the common sequence motif (given in bold, centered in the column) is upstream of.
  - **Distance:** This gives the number of bases upstream the oligomer is from the ORF associated with it in the first column. This number is the difference between the base pair number of the first base in the gene and the base pair number of the first nucleotide in the motif. It includes the distance of the promoter. This means the distance number is the difference between the promoter sequence and the ORF.
  - **Sequence:** This contains the sequence being examined written in bold. On the left side of it are the ten bases proceeding this instance of the motif, and on the right side are the 10 bases that follow it in the nucleotide sequence.

Copyright 1998-2001 Silicon Genetics

Making Lists of Homologs and Orthologs

GeneSpring's Translate feature creates a gene list in a separate genome containing genes related to genes in the current gene list. This allows you to compare genes with the same function (homologous or orthologous genes) in different organisms. In practice, however, you may choose to define any two genes in different genomes as being related.

To make lists of homologs or orthologs

1. Open the GeneSpring data folder, then open the folder of the organism you wish to translate from. Create a new folder inside this folder and name it "Homology Tables".
2. Create a text file and save it to the Homology Tables folder. In the first column of the text file, insert a unique identifier found in your master gene table for each gene in the genome you want to translate from. In the second column, insert unique identifiers for the corresponding genes from the genome you want to translate to.

   In the example below, SGD locus numbers have been used to identify genes in the yeast genome (first column), and GenBank accession numbers to identify genes in the human genome (second column).

Yeast	Human
CPR1	M80254
YDL193w	U82319
PAB1	Z48501
KGD2	D26535
YKR095w	M18533
YJL095w	U02687
YDL140c	S69370

3. Save this file with the name of the genome you are translating to and the extension *.homology.* Using the above example, this would be *Human.homology* (note that this is case sensitive).

Note that if you have a pre-4.1 version of GeneSpring you will need to take an additional step:

Open the *.genomedef* file in the folder of the genome you would like to translate to and add the following:

AcceptedDirectTranslations : Name of the genome you are translating to (without the extension)

In the above example this would be:

```
AcceptedDirectTranslations: Human
```

Copyright 1998-2001 Silicon Genetics

4. Restart GeneSpring.
5. Right-click a gene list in the genome you wish to translate from and select the Translate menu option. A submenu containing the genome you have translated to will appear. Select this option.
6. Open the genome you have translated to. You will find your new gene list in the Gene Lists folder.

Scripts

Using Scripts

New in GeneSpring 4.1 is the ability to automate complicated analyses with scripts. GeneSpring 4.1 includes several example scripts to demonstrate the power and flexibility of scripting. If you wish to design your own scripts you will need to install the Script Editor. For information on purchasing the Script Editor, please visit the Silicon Genetics Web site at http://www.sigenetics.com/Products/ScriptEditor.

To Execute a Sample Script

1. In the Navigator, open the **`Scripts > examples > high correlations folder`**.
2. Click one of the example scripts. The Run Script window will appear.
3. Choose the inputs that are required for the script by selecting a data object from the navigator panel and clicking the appropriate button in the Inputs box.
4. If the script contains knobs, you will need to enter parameters to direct the execution of the script.
5. Once all the inputs and knobs have been selected or entered, click the **`Execute locally`** button at the bottom of the window.

You can access the Script Inspector by right-clicking over any script and selecting Inspect.

Note: If you have a connection to GeNet and are using Remote Execution Servers, you have the option of having the script executed on a remote computer. To run a script remotely, do steps 1-4 as described above and click the Execute Remotely button.

What is a Script?

Scripts are tools that save time by allowing a long series of data analysis steps to be performed at once. Scripts are re-usable and can be applied to any data set. You can create your own scripts using Silicon Genetics Script Editor. All scripts, including complimentary scripts shipped with GeneSpring 4.1, are stored in the Scripts Folder.

Copyright 1998-2001 Silicon Genetics

Scripts in GeneSpring

There are seven pre-prepared scripts in the Script folder that you can use.

- **Make Gene List from Text Search:** This script will find the genes annotated with either search term 1 or search term 2 and exclude all genes with search term 3.
- **Find Similar genes:** This script will make a gene list of similar genes for every gene on the input list if there are at least 5 genes with similar expression profiles in the input experiment.
- **2-fold expression change:** This script will make a gene list of all genes that are 2-fold over-expressed or 2-fold under expressed in at least 1 condition in the input experiment.
- **Clustering 2-fold change list:** This script will make a gene tree, an experiment tree, a k-means classification, & a self organizing map using a list of all the genes that are 2-fold over-expressed or 2-fold under-expressed in at least 1 condition in the input experiment.
- **Send Clustering Results to GeNet:** This script will make a gene tree, an experiment tree, a k-means classification, & a self organizing map using a list of all the genes that are 2-fold over-expressed or 2-fold under-expressed in at least 1 condition in the input experiment and send all the results to GeNet.
- **Best k-means:** This script tries a K-means classification with 3, 5, 8 and 15 clusters, and choose the one with the highest explained variability
- **Select k-means:** This script tries 2 k-means with user input number of clusters and choose the k-means classification with the highest explained variability

Typically the scripts will divide you data into groups (such a samples or conditions) and perform analysis on these groups (sets). A group can be gene lists or conditions. Scripts create and process groups. You can create many groups, possibly more than GeneSpring can handle at one time.

The Script Inspector

Within GeneSpring you can right-click over any script and select **`Inspect`** to examine that particular script. In the Script Inspector you can edit the notes and history of your script.

Using the Remote Server

For computational intensive scripts, it is recommended you use the remote server option. This will send your data to a remote computer and allow you to keep working speedily at your local computer.

Copyright 1998-2001 Silicon Genetics

Creating Your own Scripts

The first step will be purchasing and installing the Script Editor.

Once the Script editor is installed, just click on the icon on the desktop.

There are several scripts already in your GeneSpring program. You cannot delete these scripts.

You can select the various building blocks to make a script. For a really long or intensive script, you may one to make several little scripts and them join them together.

Inputs

Inputs can go only to one place. Input will appear at the top of the screen as icon identifying lists, genome or other dataobject. Inputs will be joined from item to item by lines. these lines are thin lines for only one item, and thick lines for groups. Blue lines indicate a valid pathway, red lines indicate a possible problem. details will be given at the bottom of the screen.

Knobs

Knobs are user-defined variables. Look in the basic knobs section on the right middle of the window for drop-down menus of options (frequently the type of data to be used, see “Data Types for Restrictions” on page 4-7). This allows for greater flexibility as you can define whatever you need at the moment for the script to function.

Outputs

Multiple outputs are acceptable to GeneSpring, but if there are many new windows resulting from your script you may see a warning message before the are displayed. Outputs can be displayed in GeneSpring or saved automatically to GeNet. If there is no output in your current script there will be a warning line at the bottom of the window.

Saving your Scripts

When you are done and no more error or warning essages appear, you can save your script by clicking the **Save** button.

If you get an error message saying your result cannot be saved, rename your result and try saving again.

GeneSpring only checks for new scripts and loads them at startup, so if you make a new script in the middle of your GeneSpring Session you will need to close and re-start GeneSpring.

Copyright 1998-2001 Silicon Genetics

The Building Blocks of Scripts

Already in your script editor are various primitive building blocks you can join together in various ways to build scripts. There are several categories of building blocks.

1. Boolean
    - **Boolean:** [Generates a True or False result.]No inputs. Knob for true or false. Output is a Boolean (True or false)
    - **Boolean AND:** [Output is true if and only if both inputs are true.] 2 Boolean inputs. Output is a Boolean.
    - **Boolean False:** [Returns the result False.] No inputs. Output is a Boolean (False).
    - **Boolean NOT:** [The Boolean output is True if and only if the input is False (Converts true to false & false to true).] 1 Boolean input. Output is a Boolean.
    - **Boolean OR:** [Output is True if and only if either input is True.] 2 Boolean inputs. Output is a Boolean.
    - **Boolean True:** [Returns the result True.] No inputs. Output is a Boolean (True).
2. Boolean Select
    - **Select Boolean:** [Selects 2nd Boolean input if 1st input is true and selects 3rd Boolean input if 1st is false.] 3 Boolean inputs. Output is a Boolean.
    - **Select Condition:** [Selects 1st Condition if Boolean is True and selects 2nd Condition if Boolean is false.] 1 Boolean input & 2 condition inputs. Output is a Condition.
    - **Select Experiment:** [Selects 1st Experiment interpretation if Boolean is True and selects 2nd Experiment interpretation if Boolean is false.] 1 Boolean input & 2 Experiment interpretation inputs. Output is an Experiment interpretation.
    - **Select Experiment Tree:** [Selects 1st Experiment tree if Boolean is True and selects 2nd Experiment tree if Boolean is false.] 1 Boolean input & 2 Experiment tree inputs. Output is an Experiment tree.
    - **Select Gene:** [Selects 1st Gene if Boolean is True and selects 2nd Gene if Boolean is false.] 1 Boolean input & 2 Gene inputs. Output is a Gene.
    - **Select Gene Classification:** [Selects 1st Classification if Boolean is True and selects 2nd Classification if Boolean is false.] 1 Boolean input & 2 Classification inputs. Output is a Classification.
    - **Select Gene List:** [Selects 1st Gene List if Boolean is True and selects 2nd Gene List if Boolean is false.] 1 Boolean input & 2 Gene List inputs. Output is a Gene List.
    - **Select Gene Tree:** [Selects 1st Gene tree if Boolean is True and selects 2nd Gene tree if Boolean is false.] 1 Boolean input & 2 Gene tree inputs. Output is a Gene tree.
    - **Select Number:** [Selects 1st Number if Boolean is True and selects 2nd Number if Boolean is false.] 1 Boolean input & 2 Number inputs. Output is a Number.

Copyright 1998-2001 Silicon Genetics

- **Select Sequence:** [Selects 1st Sequence if Boolean is True and selects 2nd Sequence if Boolean is false.] 1 Boolean input & 2 Sequence inputs. Output is a Sequence.

3. Clustering
   - **Build Experiment Tree:** [Makes an Experiment Tree] 1 Gene List input & 1 Experiment interpretation input. Knobs for Correlation type, Separation ratio, & Minimum distance. Output is an Experiment Tree.
   - **Build Gene Tree:** [Makes a Gene Tree] 1 Gene List input & 1 Experiment interpretation input. Knobs for Correlation type, Discard bad, Separation ratio, Minimum distance, Do automatic annotation, & Use standard lists. Output is a Gene Tree.
   - **Explained Variation:** [Computes the proportion of variation in an experiment interpretation explained by a classification and a gene list.] 1 Classification input, 1 Experiment interpretation input, & 1 Gene List input. Output is a number between 0 & 1 inclusive. (i.e. 0.14567 is 14.567% explained variability)
   - **K-means:** [Makes a k-means classification] 1 Gene List input & 1 Experiment interpretation input. Knobs for Number of groups, Correlation type, Maximum iterations, Additional tries, & Discard bad. Output is a Classification.
   - **Refine K-means:** [Make a k-means clustering starting from a classification] 1 Classification input, 1 Gene List input & 1 Experiment interpretation input. Knobs for Correlation type, Maximum iterations, & Discard bad. Output is a Classification.
   - **Self Organizing Map:** [Makes a SOM] 1 Gene List input & 1 Experiment interpretation input. Knobs for Iterations, Discard bad, Rows, Columns, & Radius. Output is a Classification.
4. Filtering
   - **Filter Fold Change:** [Determines fold change for each gene between 2 conditions and generates a gene list with associated numbers of the genes that have a large enough fold change to pass the filter] 2 Condition inputs. Knob for Fold change. Output is a Gene List.
   - **Filter Genes with Associated Numbers:** [Takes a gene list and produces a gene list containing the genes whose associated value is above the specified parameter] Gene List input. Knobs for Cutoff & Comparison. Output is a Gene List with associated numbers.
   - **Filter On Condition:** [Produces a gene list containing the genes that have a measurement relative to a cutoff] 1 Condition input. Knobs for Filter type, Filter cutoff, & Comparison. Output is a Gene list.
   - **Filter on Gene Correlation:** [Find the genes that have a certain correlation in an experiment (Find Similar Genes)] 1 Gene input & 1 Experiment interpretation input. Knobs for Correlation type, Cutoff, & Comparison. Output is a Gene List with associated numbers.
   - **Filter on Text in Description:** [Find genes containing the specified text] 1 Gene list input. Knob for Search term. Output is a Gene List.

Copyright 1998-2001 Silicon Genetics

5. Gene List Manipulation

- **All Genes:** [Result is All Genes list.] No inputs or knobs. Output is All Genes Gene List.
- **All Genomic:** [Result is All Genomic Elements list.] No inputs or knobs. Output is All Genomic Elements Gene List.
- **Gene List Difference:** [Make a Gene List of the genes that are in the first gene list, but not the second gene list.] 2 Gene List inputs. Output is a Gene List.
- **Gene List Intersection:** [Make a Gene List of the genes that are in both input gene lists.] 2 Gene List inputs. Output is a Gene List.
- **Gene List Union:** [Make a Gene List of the genes that are in either input gene list.] 2 Gene List inputs. Output is a Gene List.
- **In all Gene lists:** [Make a Gene List of the genes in all the input gene lists.] 1 Gene List Group input. Output is a Gene List.
- **In at least one:** [Make a Gene List of the genes in at least one of the input gene lists.] 1 Gene List Group input. Output is a Gene List.
- **Merge Gene List Group:** [Make a Gene List of the genes in a certain proportion (specified by knobs) of the input gene lists.] 1 Gene List Group input. Knobs for Percentage & Comparison. Output is a Gene List.
- **Number of Genes:** [Produce the number of genes in the gene list.] 1 Gene List input. Output is a number (number of genes in the gene list).

6. GeNet Publishing

a. Default Directory

- **Send Classification to GeNet:** [Publish a classification to your default directory in GeNet.] 1 Classification input. (No knobs or outputs.)
- **Send Experiment to GeNet:** [Publish an Experiment interpretation to your default directory in GeNet.] 1 Experiment interpretation input. (No knobs or outputs.)
- **Send Experiment Tree to GeNet:** [Publish an Experiment tree to your default directory in GeNet.] 1 Experiment Tree input. (No knobs or outputs.)
- **Send Gene List to GeNet:** [Publish a Gene List to your default directory in GeNet.] 1 Gene List input. (No knobs or outputs.)
- **Send Gene Tree to GeNet:** [Publish a Gene Tree to your default directory in GeNet.] 1 Gene Tree input. (No knobs or outputs.)

b. Specified Directory

- **Send Classification to Directory in GeNet:** [Publish a classification to a chosen directory in GeNet.] 1 Classification input. Knob for Directory. (No outputs.)

Copyright 1998-2001 Silicon Genetics

- **Send Experiment to Directory in GeNet:** [Publish an Experiment interpretation to a chosen directory in GeNet.] 1 Experiment interpretation input. Knob for Directory. (No outputs.)
- **Send Experiment Tree to Directory in GeNet:** [Publish an Experiment tree to a chosen directory in GeNet.] 1 Experiment Tree input. Knob for Directory. (No outputs.)
- **Send Gene List to Directory in GeNet:** [Publish a Gene List to a chosen directory in GeNet.] 1 Gene List input. Knob for Directory. (No outputs.)
- **Send Gene Tree to Directory in GeNet:** [Publish a Gene Tree to a chosen directory in GeNet.] 1 Gene Tree input. Knob for Directory. (No outputs.)

7. Groups
   - **Merge Genes:** [Merges a group of genes into a gene list.] 1 Gene Group input. Output is a Gene List.
   - **Merge Genes and Numbers:** [Merges a group of genes into a gene list with associated numbers. If the genes and numbers do not match, the results are undefined.] 1 Gene Group input & 1 Number group input. Output is a Gene List.
   - **Split Classification:** [Splits the classification up into a group of gene lists.] 1 Classification input. Output is a Group of Gene Lists.
   - **Split Conditions:** [Splits the Experiment interpretation into a group of Conditions.] 1 Experiment interpretation input. Output is a Group of Conditions.
   - **Split Gene List:** [Splits the Gene List up into a Group of Genes.] 1 Gene List input. Output is a Group of Genes.
   - **Split Gene List With Numbers:** [Splits the Gene List up into a Group of Genes and an associated Group of Numbers.] 1 Gene List input. Output is a Group of Genes & a Group on Numbers.
8. Filter
   - **Filter Boolean Group:** [For each Boolean in the first argument, pass through the corresponding second argument if the Boolean is true.] 2 Boolean Group inputs. Output is a Boolean Group.
   - **Filter Condition Group:** [For each Boolean in the first argument, pass through the corresponding Condition if the Boolean is true.] 1 Boolean Group input & 1 Condition Group input. Output is a Group of Conditions.
   - **Filter Experiment Group:** [For each Boolean in the first argument, pass through the corresponding Experiment interpretation if the Boolean is true.] 1 Boolean Group input & 1 Experiment interpretation Group input. Output is a Group of Experiment interpretations.
   - **Filter Experiment Tree Group:** [For each Boolean in the first argument, pass through the corresponding Experiment Tree if the Boolean is true.] 1 Boolean Group input & 1 Experiment Tree Group input. Output is a Group of Experiment Trees.

Copyright 1998-2001 Silicon Genetics

- **Filter Gene Group:** [For each Boolean in the first argument, pass through the corresponding Gene if the Boolean is true.] 1 Boolean Group input & 1 Gene Group input. Output is a Group of Genes.
- **Filter Gene Classification:** [For each Boolean in the first argument, pass through the corresponding Classification if the Boolean is true.] 1 Boolean Group input & 1 Classification Group input. Output is a Group of Classifications.
- **Filter Gene List Group:** [For each Boolean in the first argument, pass through the corresponding Gene List if the Boolean is true.] 1 Boolean Group input & 1 Gene List Group input. Output is a Group of Gene Lists.
- **Filter Gene Tree Group:** [For each Boolean in the first argument, pass through the corresponding Gene Tree if the Boolean is true.] 1 Boolean Group input & 1 Gene Tree Group input. Output is a Group of Gene Trees.
- **Filter Number Group:** [For each Boolean in the first argument, pass through the corresponding Number if the Boolean is true.] 1 Boolean Group input & 1 Number Group input. Output is a Group of Numbers.
- **Filter Sequence Group:** [For each Boolean in the first argument, pass through the corresponding Sequence if the Boolean is true.] 1 Boolean Group input & 1 Sequence Group input. Output is a Group of Sequences.

9. Look Up
    - **Number associated with gene in Condition:** [Return the number (0 if none) associated with a gene in a condition.] 1 Gene input & 1 Condition input. Knob for Type. Output is a Number.
    - **Number associated with gene in Gene List:** [Return the number (0 if none) associated with a Gene in a Gene List.] 1 Gene input & 1 Gene List input. Output is a Number.
    - **See if Gene List contains a gene:** [Return True if a Gene List contains a given Gene.] 1 Gene input & 1 Gene List input. Output is a Boolean.
10. Numbers
    - **Compare 1 number:** [Compare a number to another number specified as a parameter.] 1 Number input. Knobs for Comparison & Number. Output is a Boolean.
    - **Compare 2 numbers:** [Compares two numbers.] 2 Number inputs. Knob for Comparison. Output is a Boolean.
    - **Number:** [Produce the number specified in the parameter.] Knob for Number. Output is a Number.
    - **Number Add:** [Add two numbers together.] 2 Number inputs. Output is a number.
    - **Number Div:** [Divide the first number by the second number.] 2 Number inputs. Output is a number.
    - **Number Mul:** [Multiply two numbers together.] 2 Number inputs. Output is a number.

Copyright 1998-2001 Silicon Genetics

- **Number Sub:** [Subtract the second number from the first number.] 2 Number inputs. Output is a number.

11. Promoter

- **Find Genes in GeneList with Regulatory Sequence:** [Produces a Gene List showing the genes that contain the input regulatory Sequence.] 1 Sequence input & 1 Gene List input. Knobs for From Base, To Base, & Maximum errors. Output is a Gene List.
- **Find Genes with Regulatory Sequence:** [Produces a Gene List showing the genes that contain the input regulatory Sequence.] 1 Sequence input. Knobs for From Base, To Base, & Maximum errors. Output is a Gene List.
- **Find Regulatory Sequence:** [Find regulatory sequences upstream of the genes in the Gene List specified as input.] 1 Gene List input. Knobs for From Base, To Base, Minimum Length, Maximum Length, Minimum Errors, Maximum Errors, Minimum Interior N's, Maximum Interior N's, Relative Genomic, p-value cutoff. Output is a Group of Sequences.

Auto-Publish to GeNet

You can also use Scripts to automate publishing to GeNet.

External Programs

GeneSpring External Program Interface

The GeneSpring™ External Program interface allows you to run external analysis programs from within GeneSpring. These programs can be useful when your research calls for a type of analysis that GeneSpring does not perform. The external program interface is also useful for parsing and pre-formatting data for use in another application.

When you launch an external program from within GeneSpring, the data that is displayed in the genome browser will be sent to the external program as standard input. When the external program runs, GeneSpring recognizes the standard output generated by the external program and displays it in the genome browser.

To run an External Program

1. Select the gene list that you want to send to the program.
2. If your program takes the data from a tree or a classification as input, be sure these are selected and visible as well.
3. Open the external program folder in the navigator panel and click the program you wish to run.

Copyright 1998-2001 Silicon Genetics

To install a new external program

1. Create or obtain an external program. Any program capable of receiving standard input is acceptable.
2. Create a file named XXXXXX.programdef. Each line of a *.programdef* file should contain a parameter, followed by a colon, followed by the parameter value. Blank lines and lines beginning with the `#' sign will be ignored. GeneSpring recognizes the following parameters.

   - **Name (required):** the name of the external program as it will appear in the navigator. For example:
     `Name : Sort Gene List`
   - **Icon (optional):** the file name of a 16x16 pixel *.gif* file that includes an icon to be displayed in the navigator. For example:
     `Icon : sorter.gif`
   - **Command (required):** the command line string required to run the program. For example:
     `Command : Sort`
     or
     `Command : perl sort.pl`
   - **Input (required):** one or more numbers separated by commas corresponding to the type(s) of input that the external program requires (see table XXX). For example:
     `Input : 2, 5`
   - **Output (required):** one or more numbers corresponding to the type of output that the external program sends to GeneSpring (see table XXX [Include existing table at the end of this section]). For example:
     `Output : 2`
   - **UserParameters (optional):** one or more user-defined parameters separated by commas that are passed to the external program. For example:
     `UserParameters : Iterations=10000`
   - **UserParameterFill (optional):** a text string to fill in blank values for the UserParameters above. For example:
     `UserParameterFill : none`
   - **GeneListNumberDescription (optional):** if the external program returns an ordered gene list back to GeneSpring. For example:
     `GeneListNumberDescription :`
   - **TerminateWith255:** true if you want GeneSpring to terminate the external program input with ASCII 255. For example:
     `TerminateWith255 : true`
   - **InterModeDelimiter (optional):** an ASCII code representing the character used to delimit multiple objects that are sent to the external program. For example:
     `InterModeDelimiter : 255`

Copyright 1998-2001 Silicon Genetics

- **DebugInput (optional):** true if you want the data that is passed to the external program to be displayed in the Java console. For example:
  `DebugInput : true`
- **DebugOutput (optional):** true if you want the data that is passed from the external program back to GeneSpring to be displayed in the Java Console. For example:
  `DebugOutput : true`

3. Place the *.programdef* file in the Programs folder in your GeneSpring/Data directory.

Examples

External Program Interface Example #1: SAS™ for Windows

This example demonstrates how to use GeneSpring's external program interface. The External Program Interface will export GeneSpring experimental data, run a SAS™ program to analyze it, and bring the results back into GeneSpring for display. This example has been developed with Windows 2000, but should work with earlier versions of Windows. It uses SAS™ Version 8, and you will need to change it somewhat to work with earlier versions of SAS™.

This particular example sets up an interface to the SAS™ procedure FASTCLUS to do gene clustering. You will need to create three text files with a text editor such as Microsoft NotePad™. These files are FASTCLUS.programdef, Runsas.bat, and Fastclus.sas. These are each described below. The first line of the description gives the name of the file (including the proper file extension), and the location where the file should be placed. The file placement relies upon having the default directory set to ...\GeneSpring\data as part of the GeneSpring setup. This allows you to avoid having to write out the full path names of the Runsas.bat, and Fastclus.sas files within FASTCLUS.programdef (as long as they are placed in the ...\GeneSpring\data directory). The *.programdef* file, must be in the Programs subfolder of ...\GeneSpring\data directory. If you don't already have a Programs subfolder in this directory, create one. The code following the title and location of the file should be entered as the text of that file.

In the ...\GeneSpring\data\Programs put this file:

...\GeneSpring\data\Programs\FASTCLUS.programdef

```
# External Program interface for SAS
Name: FASTCLUS
Command: runsas.bat fastclus expt.txt clus.txt
Input: 4
Output: 6
```

This file defines four things (see the External Program Interface FAQ for details.):

- The displayed name in GeneSpring
- The input format for the experimental data going into SAS™
- The output format for the cluster membership data coming back from SAS™
- The name of the batch file actually doing the work.

Copyright 1998-2001 Silicon Genetics

In the ...\GeneSpring\data directory place these two files:

...\GeneSpring\data\Runsas.bat

```
@echo off
set infile=%2
set outfile=%3
cat.exe > %2
C:\PROGRA~1\SASINS~1\SAS\V8\SAS.EXE %1.sas -nologo -config +
C:\PROGRA~1\SASINS~1\SAS\V8\SASV8.CFG cat.exe < %3
del %1.lst %1.log %2 %3
```

(Note: When you are preparing this file, remove the plus sign and combine the two lines beginning with C:\PROGRA~1 into one long line.)

This batch file takes the standard input from GeneSpring, stores it in a file, executes SAS™, and then passes the results back to GeneSpring via standard output. The program cat.exe simply copies standard input into standard output, if you do not have something equivalent on your system, cat.exe can be downloaded from Silicon Genetics' web site.

...\GeneSpring\data\Fastclus.sas

```
filename infile "%sysget(infile)";
filename outfile "%sysget(outfile)";
proc import datafile=infile DBMS=TAB out=experiment replace;
datarow=3;
getnames=no;
run;
proc fastclus data=experiment maxclusters=5 maxiter=50
out=clusters(keep=var1 cluster);
id var1;
run;
proc export data=clusters outfile=outfile DBMS=TAB replace;
run;
```

This runs PROC FASTCLUS, specifying 5 clusters. In PROC IMPORT, the datarow=3 command skips the first 2 lines of the exported data, which contain the dataset name and one parameter. If you have more than one parameter, you should adjust the data-row value accordingly.

PROC EXPORT puts a header line on the return data set listing the variable names, and GeneSpring will give you an error message and should skip this line (unless you have a gene named VAR1, in which case you should rename VAR1 to something else in your application).

Once you have all three files set up, restart GeneSpring, and open the External Programs folder. There should be an entry named FASTCLUS. If you select this item, you will see SAS™ put up a batch window while it is running, then GeneSpring will come back with a classification based on the SAS™ clustering, and you can save and work with the classification in GeneSpring.

Copyright 1998-2001 Silicon Genetics

Example - File Access

The File Access external programs are a set of Java programs written using the GeneSpring External Program Interface that allow you to read and write GeneSpring data objects to and from files. These functions are:

```
Load Classification From File
Load Experiment From File
Load Gene List From File
Load Gene List With Numbers From File
Load Tree From File

Save Classification To File
Save Experiment To File
Save Gene List To File
Save Gene List With Numbers To File
Save Tree To File
```

These correspond to the data formats previously discussed (*Experiment* here means Experiment Data with Confidence). These provide convenient alternatives to using the clipboard to copy and paste data from GeneSpring. To use the Save features, select the object you wish to export, and then click on the corresponding Save command. A file naming dialog will appear to allow you to name the output file. To use the Load feature, click on the appropriate Load command, a file selector dialog will appear to allow you to choose the file to load, and when the data is loaded, then a new data object dialog will appear to allow you to name the data object, and put it in a GeneSpring folder if you desire.

These programs are all contained in one jar file called *FileAccess.jar*, that needs to be placed in the Programs subfolder of the GeneSpring Data folder on your hard disk. You can get the latest version of this file from

http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/extProgs.smf

Download the jar, create a Programs folder in your GeneSpring Data folder (if needed), put the jar file in it, and restart GeneSpring. You should now have several new items under the External Programs menu in the GeneSpring navigator. If your External Programs menu is getting cluttered, you can create a folder within the Programs folder (e.g. *File Access*) and put the *FileAccess.jar* file inside that folder, the File Access items will then appear in the correspondingly named subfolder of the External Programs folder.

Copyright 1998-2001 Silicon Genetics

Chapter 5 Clustering and Characterizing Data in GeneSpring

Trees

The classification of organisms into phylogenetic trees is a central concept to biology. Organisms sharing properties tend to be clustered together. How far up the tree you have to go to find a branch containing both organisms can be considered a measure of how different the organisms are. You can classify genes in a similar manner—clustering those whose expression patterns are similar into nearby places in a tree. Such mock-phylogenetic trees are often referred to as dendrograms.

GeneSpring can both create and display such trees. GeneSpring can also create trees of experiments, displaying the genes along the X-axis and the samples along the Y-axis. This can be exceedingly powerful for many applications; for example, seeing if any environmental stressors cause similar effects on the expression levels as mutant organisms do.

If you have already created or downloaded trees, open the Gene Trees folder in the navigator and select any tree for viewing.

Creating a New Gene Tree

For detailed instructions on creating a Gene Tree in GeneSpring with the default values, please refer to *GeneSpring Basics Instructional Manual* Chapter 6 “Trees” on page 6-1.

While viewing any list:

1. In the main GeneSpring screen, select **`Tools > Clustering`**.
2. In the Clustering window, select **`Make New Tree`** from the drop-down list labeled Clustering Method.
3. Select the **`Start`** button at the bottom of the screen. This will start the process of computing and annotating a gene tree. As this is a computationally intensive process, it could take a few minutes. A Clustering Progress bar will indicate the progress of the clustering.

   Clicking the **`Start`** button will not close the Clustering window, so you can begin planning another tree immediately. For details on all the options you could change, please refer to “Creating Complex Experiment Trees” on page 5-2. Changing the information given in the Clustering window after you have started clustering a tree does not change the parameters of the tree in the process of being made. Changing the parameters displayed changes the parameters required for the next tree you make from this window. The Close button, at the bottom of the window, closes the Clustering window. This will not halt the making of a tree currently in the process of clustering. You cannot start clustering a new tree while there is already one in the process of being computed.

Copyright 1998-2001 Silicon Genetics

4. The Name New Tree window will appear. Name your tree and select **Save**.
5. GeneSpring will automatically take you back to the main window where you can examine your new tree. You may need to resize the window by clicking and dragging the edges in order to view the parameters.

You can also view another list in this same tree structure by selecting a new list from the Gene Lists folder.

Creating Complex Experiment Trees

Complex trees can be made from multiple experiments or by tightly defining the types of data to use. You can select a gene list the navigator to reduce the number of genes to be made into a tree.

To begin an Experiment Tree

1. Select **Tools > Clustering**.
2. Select **Experiment Tree** from the Clustering Method pull-down menu.
3. Select a gene list from the Gene Lists folder in the Clustering window.
4. To add an experiment, interpretation or condition, click on one of these items in the Experiments folder of the Clustering window, click the **Add** button in the Experiments to Use section and enter a weight in the pop-up window.

   Or,

   Right-click an experiment or condition in the Clustering window ad choose **Add Experiment Correlation** from the pop-up menu. Enter a weight in the pop-up menu and click **OK**.

   - You can add multiple experiments, interpretations or conditions.
   - You can right-click experiment, interpretation or condition to add a restriction. See “Filter Genes Analysis Tools” on page 4-1 and “Making Lists with the Complex Correlation Command” on page 4-14 for details.
5. Choose a measure of similarity from the pull-down menu. See “Equations for Correlations and other Similarity Measures” on page L-1 for details.
6. Choose a separation ratio. See “Minimum Distance and Separation Ratios” on page 5-3.
7. Choose a minimum distance. See “Minimum Distance and Separation Ratios” on page 5-3.
8. Click **Start**.

   Note: You can right-click the list to Add Associated Numbers Restriction if desired. See “Adding an Associated Number Restriction” on page 4-9.

Correlations of multiple experiments are done through a weighted correlation, in which you specify the weight of each experiment. You may make one experiment or experiment set more important than another. If all of the experiments, or experiment sets, are given the same weight, they will be averaged equally. The name of the experiment is noted directly after its relative weight. For example, you could give SampleExperiment1 a weight of 2, and Experiment2 a weight of 1.

 Copyright 1998-2001 Silicon Genetics

Therefore, in this example, the correlations found in the SampleExperiment1 will be twice as influential in creating the tree as the correlations between the genes in the Experiment2 study.

The equation used to determine the overall correlation is:

$$X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}$$

- **A** is the correlation coefficient between the gene in question in experiment 1 and the gene named in the *Experiments to Use* box, also from experiment 1.
- **a** is the weight specified for experiment 1.
- **B** is the correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- **b** is the weight associated with experiment 2.
- **C** is the correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- **c** is the weight associated with experiment 3.

and so on.

Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white *Correlations* box. If **X** is between the minimum and maximum correlations specified in the Clustering window, then the gene in question passes the correlations.

To Delete an Experiment from the Current Clustering

1. Click the name of the experiment in the white *Experiments to Use* window, highlighting it.
2. Click the **Remove** button.

Similarity Definitions

The equations used to determine the nine types of correlations are described in detail in "Equations for Correlations and other Similarity Measures" on page L-1.

The default correlation is the Standard Correlation, *Standard correlation* = $\mathbf{a}.\mathbf{b}/(|\mathbf{a}||\mathbf{b}|)$.

Minimum Distance and Separation Ratios

To make a tree, GeneSpring calculates the correlation for each gene with every other gene in the set. Then it takes the highest correlation and pairs those two genes, averaging their expression profiles. GeneSpring then compares this new composite gene with all of the other unpaired genes. This is repeated until all of the genes have been paired. At this point the minimum distance and the separation ratio come in to play. Both of these affect the branching behavior of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A value smaller than .001 has very little effect, because most genes are not correlated more closely than that. A higher number will tend to lump more genes into a group, making the groups less specific. The separation ratio determines how large the correlation difference between groups of clustered genes has to be for them to be considered discrete groups, and not be lumped together. This number should be between 0 and 1.

 Copyright 1998-2001 Silicon Genetics

It is not normally appropriate to change separation ratio or minimum distance.

- Separation Ratio

  The separation ratio determines how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

  - Increasing separation increases the ‘branchiness’ of the tree.
  - Default Separation ratio is 0.5. Separation ratio can range from 0.0 to 1.0.
  - At a separation ratio of 0, all gene expression profiles can be regarded as identical.

  To change the maximum correlation number highlight the number in the white box next to the *Separation Ratio* label, and type in a new value. You will not normally want to modify value.

- Minimum Distance

  The number specified in the *Minimum distance* box determines the minimum separation considered significant between genes. This reduces meaningless structure at the base of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A higher number will tend to lump more genes into a group, making the groups less specific.

  - Decreasing minimum distance increases the ‘branchiness’ of the tree.
  - Default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely.

  To change default minimum distance number move the cursor into the white box next to the *Minimum distance* label, and click in the box, then use the keyboard to alter the text, just like using a word processing program. You will not normally want to modify the minimum distance.

References for Hierarchical Clustering

Everitt, Brian S. *Cluster Analysis* (3rd Ed.) Arnold, London, 1993, pp 62-65.

Eisen, Michael B., et. al. “Cluster analysis and display of genome-wide expression patterns” *Proc. Natl. Acad. Sci. USA*, V95, pp 14863-14868, December 1998.

Copyright 1998-2001 Silicon Genetics

Principal Components Analysis

Principal components analysis (PCA) is a decomposition technique that produces a set expression patterns known as principal components. Linear combinations of these patterns can be assembled to represent the behavior of all of the genes in a given data set. It should be noted that PCA is not a clustering technique. Rather, it is a tool to characterize the most abundant themes or building blocks that reoccur in many genes in your experiment.

To perform a PCA analysis, select **`Tools > Principal Components Analysis`**.

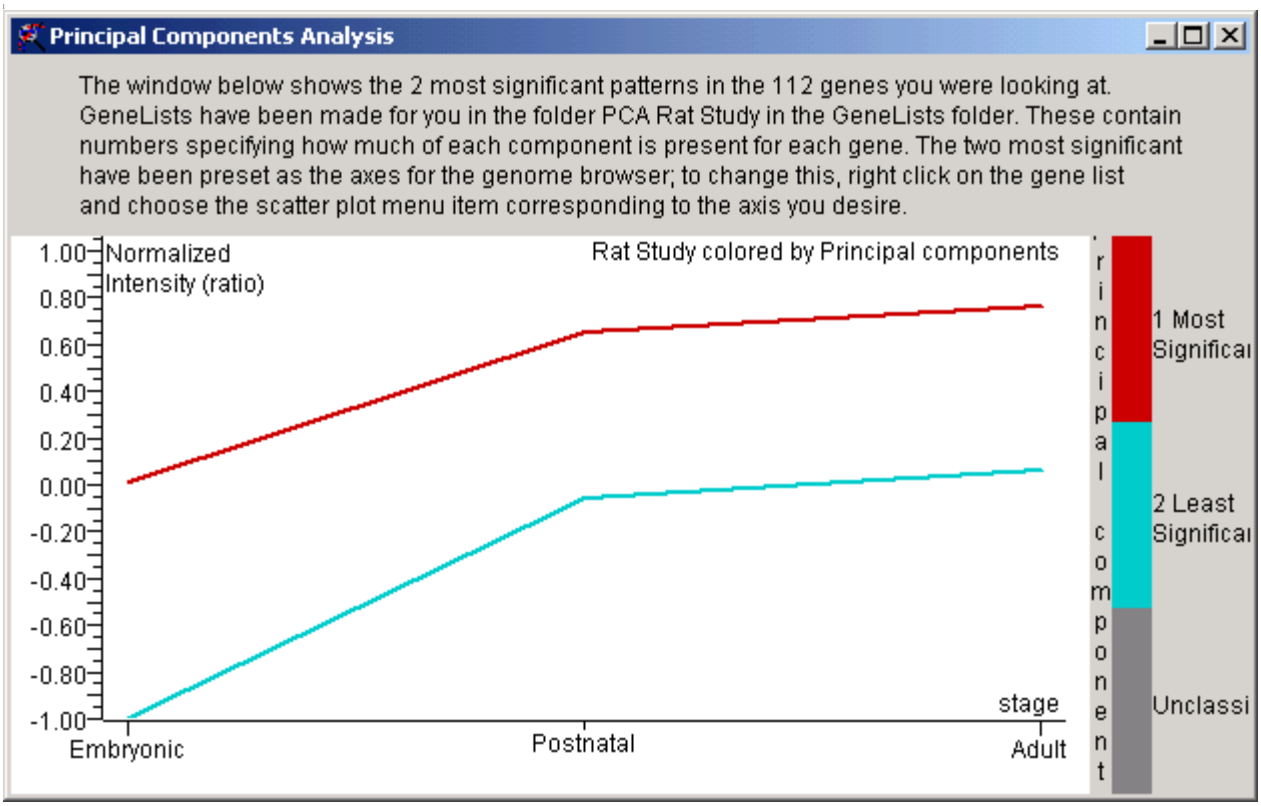

Principal Components Analysis
The window below shows the 2 most significant patterns in the 112 genes you were looking at. GeneLists have been made for you in the folder PCA Rat Study in the GeneLists folder. These contain numbers specifying how much of each component is present for each gene. The two most significant have been preset as the axes for the genome browser; to change this, right click on the gene list and choose the scatter plot menu item corresponding to the axis you desire.
Normalized Intensity (ratio)
Rat Study colored by Principal components
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-1.00
Embryonic
Postnatal
Adult
stage
Principal component
1 Most Significa
2 Least Significa
Unclassi

Figure 5-1 Principal Components Analysis window

When the analysis finishes, the Principal Components Analysis window appears, displaying each component as a line in graph mode. The significance of each component is represented by the color of its graph line, as defined by the colorbar. Double-clicking any of the components will bring up the Gene Inspector window, which shows the eigenvalue and explained variability in the upper-left panel. In addition, a new gene list folder will appear in the navigator panel with a name that includes the name of experiment that you used for PCA analysis (e.g., “PCA yeast cell cycle”).

Interpreting your PCA Results

The principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the covariance matrix of the data. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector. The eigenvector of

Copyright 1998-2001 Silicon Genetics

the largest eigenvalue is the first principal component. The eigenvector of the second largest eigenvalue is the second principal component and so on. Principal components which explain significant variability are displayed by GeneSpring in the Principal Components Analysis window.

There will never be more principal components than there are conditions in the data.

Viewing Principal Components in a Scatter Plot

After performing principal components analysis, the genome browser displays a scatter plot in which the first and second principal components (representing the largest fraction of the overall variability) are plotted on the vertical and horizontal axis respectively. This type of view is useful for selecting and making lists of genes that exhibit high levels one or two principle components. Genes that exhibit high levels of the first principal component and low levels of the second principal component are displayed in the lower right corner of the plot, and genes exhibiting equal levels of the two components lie along the diagonal.

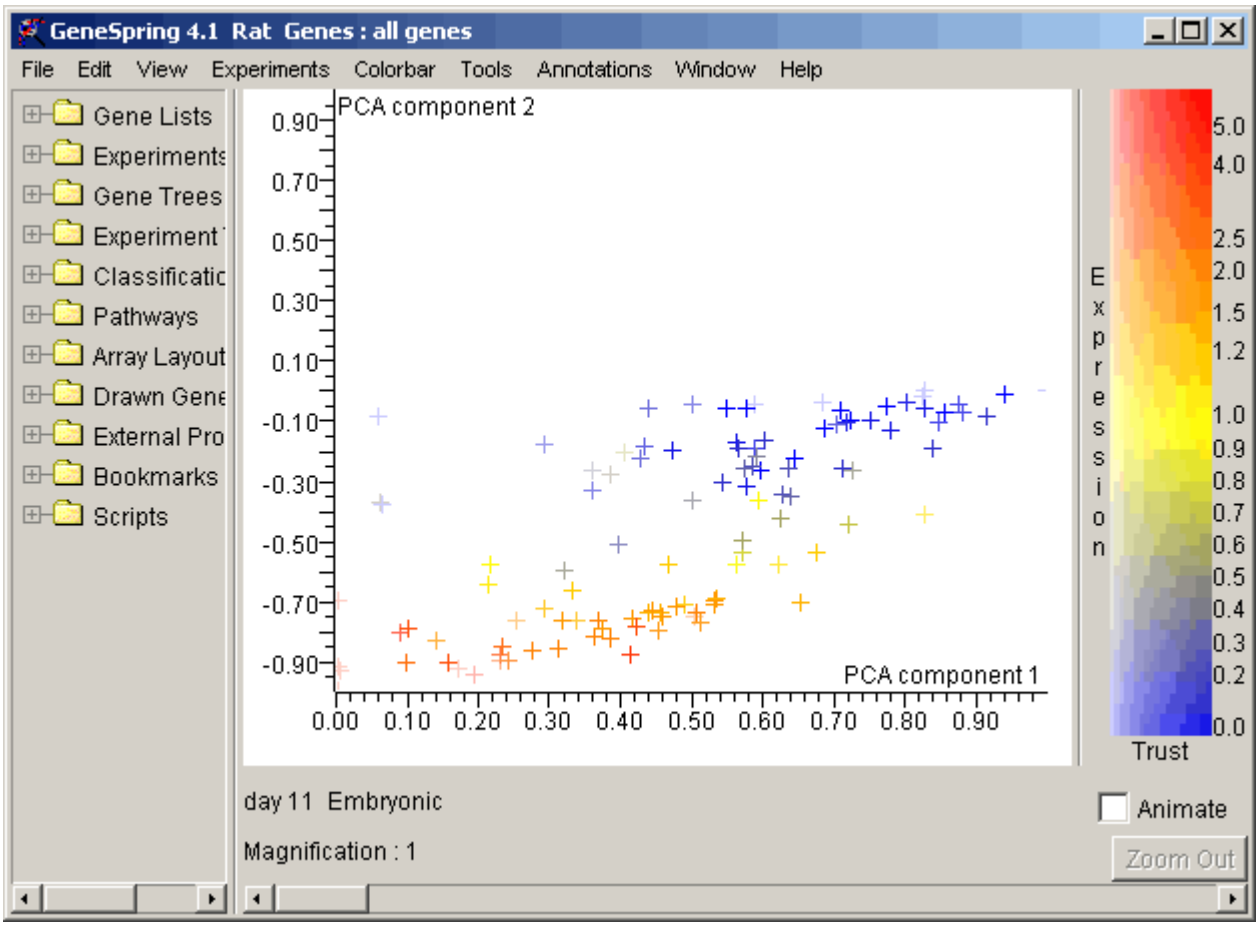


Figure 5-2 PCA Scatter Plot in Log Mode

You can change the components that are represented by each axis by right-clicking one of the gene lists in the PCA gene list folder.

Copyright 1998-2001 Silicon Genetics

Viewing Principal Components in an Ordered List

Perhaps the best way to visualize the genes that exhibit the highest levels of an individual component is to use the ordered list view. Select **`View > Ordered List`** and select one of the PCA gene lists from the navigator panel. Genes exhibiting the highest levels of the selected principal component will be displayed on the left side of the genome browser and will have the longest lines extending upward from them. For more details, please see “Ordered List View” on page 3-21.

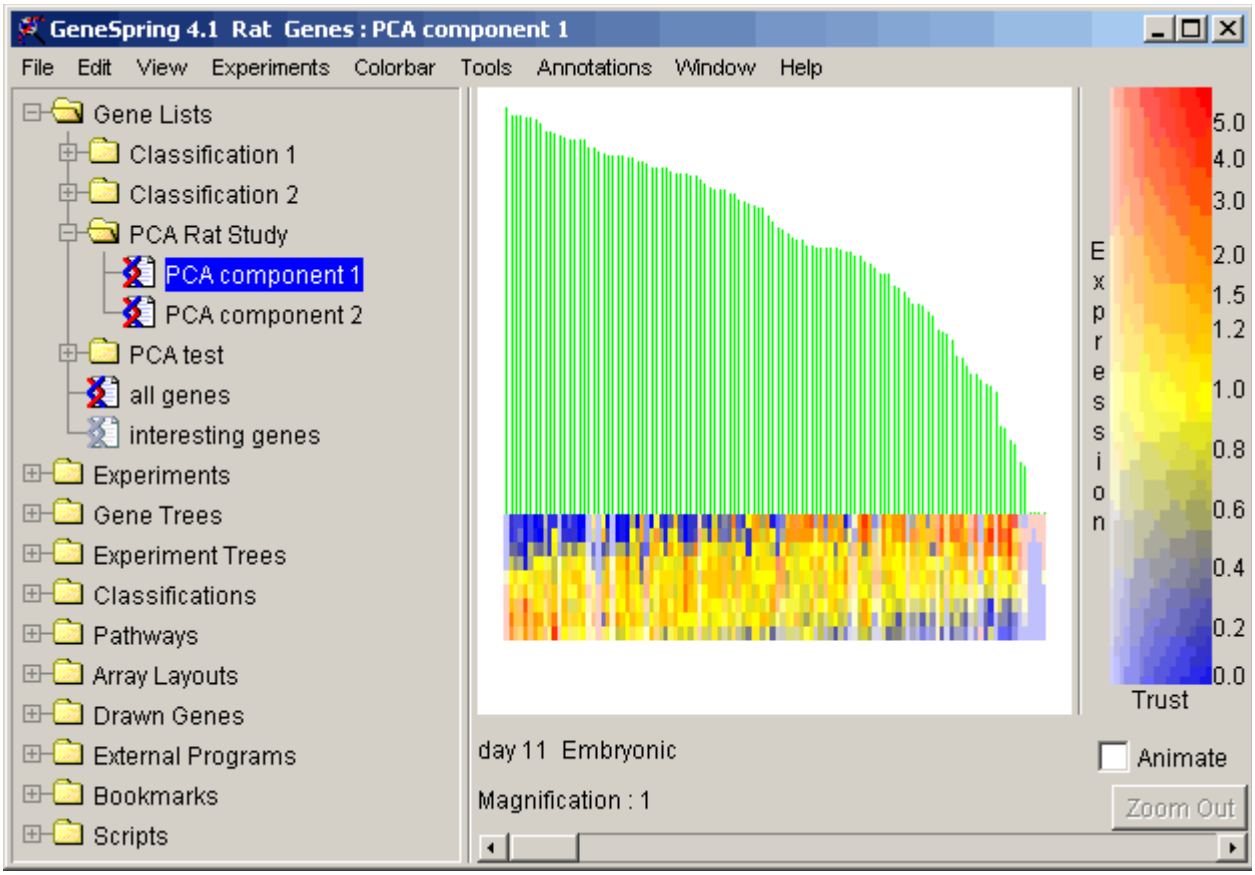


Figure 5-3 PCA in the Ordered List view

Copyright 1998-2001 Silicon Genetics

References for Principal Components Analysis

Alter O., Brown P.O., Botstein D. *Singular value decomposition for genome-wide expression data processing and modeling.* PNAS 97:10101-6 (2000) http://www.pnas.org/cgi/content/full/97/18/10101

Cooley, W.W. and Lohnes, P.R. *Multivariate Data Analysis* (John Wiley & Sons, Inc., New York, 1971).

Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations* (John Wiley & Sons, Inc., New York, 1977).

Neal S. Holter et al, Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS 97,8409 (2000) http://www.pnas.org/cgi/content/abstract/97/15/8409

Hotelling, H. *Analysis of a Complex of Statistical Variables into Principal Components*. Journal of Educational Psychology **24**, 417-441, 498-520 (1933).

Kshirsagar, A.M. *Multivariate Analysis* (Marcel Dekker, Inc., New York, 1972).

Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate Analysis* (Academic Press, London, 1979).

Morrison, D.F. *Multivariate Statistical Methods*, Second Edition (McGraw-Hill Book Co., New York, 1976).

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **6(2)**, 559 -572 (1901).

Rao, C.R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya* A **26**, 329 –358 (1964).

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pacific Symposium on Biocomputing (2000).

Copyright 1998-2001 Silicon Genetics

k-Means Clustering

k-means clustering divides genes into groups based on their expression patterns. The goal is to produce groups of genes with a high degree of similarity within each group and a low degree of similarity between groups. Unlike self-organizing maps, k-means clustering is not designed to show the relationship between clusters. Instead, k-means clusters are constructed so that the average behavior in each group is distinct from any of the other groups. For example, in a time series experiment you could use k-means clustering to identify unique classes of genes that are upregulated or downregulated in a time dependent manner.

GeneSpring's k-means clustering algorithm divides genes into a user-defined number (k) of equal-sized groups, based on the order in the selected gene list. It then creates centroids (in expression space) at the average location of each group of genes. With each iteration, genes are reassigned to the group with the closest centroid. After all of the genes have been reassigned, the location of the centroids is recalculated and the process is repeated until the maximum number of iterations has been reached.

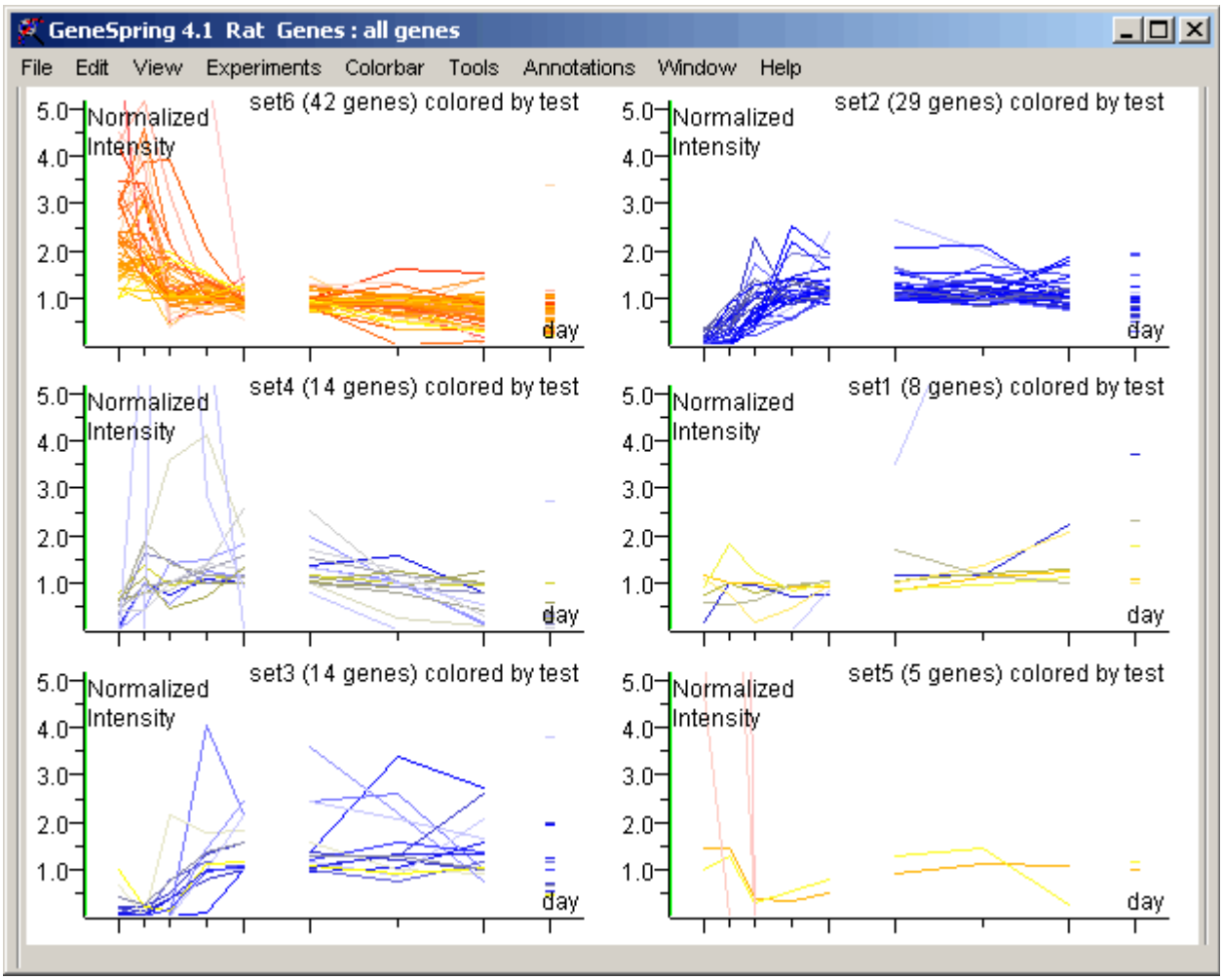

GeneSpring 4.1 Rat Genes : all genes
File Edit View Experiments Colorbar Tools Annotations Window Help
set6 (42 genes) colored by test
set2 (29 genes) colored by test
set4 (14 genes) colored by test
set1 (8 genes) colored by test
set3 (14 genes) colored by test
set5 (5 genes) colored by test
Normalized Intensity
5.0
4.0
3.0
2.0
1.0
day

Figure 5-4 A k-means Cluster display in a Split Window

Copyright 1998-2001 Silicon Genetics

To Perform k-means Clustering

1. Select **`Tools > Clustering`**. The Clustering window will appear as in Figure 5-5.

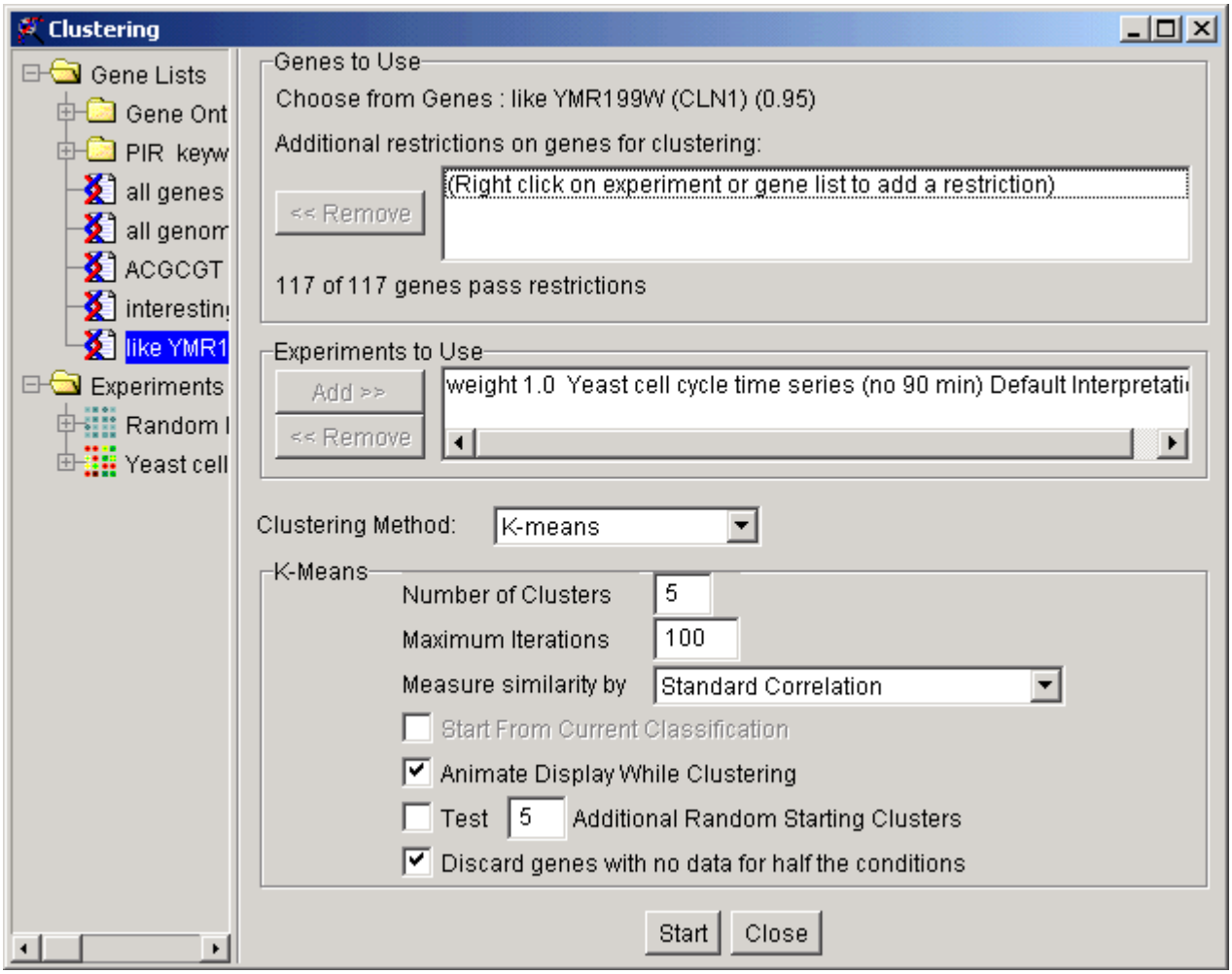


Figure 5-5 The GeneSpring Clustering window

2. Choose a gene list from the Gene List folder in the navigator, right-click the list and select **`Set Gene List`**. To remove a gene list, select the list in the Genes to Use box and click **`Remove`**.
   - To add restrictions to the selected list, right-click an experiment or gene list in the navigator and select a restriction. For information on restrictions and how to apply them, see "Filtering Genes" on page 4-1.
   - Selecting **`Discard Genes With No Data For Half The Conditions`** discards any genes with no data in at least half the conditions in the selected experiment.
3. To add an experiment or condition, click on an experiment or condition in the Experiments folder of the navigator. Enter a weight in the pop-up window. Click the **Add** button under *Experiments to Use*. To remove an experiment or condition, select the experiment or condition under *Experiments to Use* and click **`Remove`**.

Copyright 1998-2001 Silicon Genetics

- The weight of the condition is a measure of the influence the condition has on the correlation distance, e.g. an experiment with a weight of 2.0 will be twice as influential as one with a weight of 1.0.

4. Enter the **Number of Clusters** that you wish to make.
5. Choose the maximum number of iterations. This is the maximum number of times that each centroid is recalculated after genes are reassigned to groups with the most similar centroids.
6. Choose a measure of similarity. For information on measures of similarity, see "Equations for Correlations and other Similarity Measures" on page L-1. If you do not want to base the initial grouping of genes on the order of the current gene list, you can choose one of these two options for selecting starting classifications:
   - The **Start From Current Classification** feature groups genes according to the selected classification. Note that this option is only available if you have selected a classification. This option disables the **Number of Clusters** checkbox as it automatically uses the number of classes in the current classification.
   - The **Test Additional Random Starting Clusters** feature makes clustering as tight as possible by performing clustering several times, each time starting from a different random grouping of genes, and choosing the best result.
7. If you want to watch the k-means clustering process as it occurs, the **Animate Display While Clustering** feature shows changes in classification assignments in real time. This may slow your analysis slightly.
8. Click **Start**. Clustering may take a few moments depending on how many genes are being clustered and how many iterations you chose. When the clustering finishes, the Choose Classification Name window will appear.
9. Despite the name of the window, you can save the result either as a classification or as gene lists by selecting one of the two **Save Classification as:** radiobuttons. Select a name for your classification/list and click **Save**.

Viewing k-means clusters

If you use k-means clustering to produce a classification, you can get details about the classification in the Classification Inspector. For information about the Classification Inspector, see "Classification Inspector" on page 3-46.

Perhaps the easiest way to view a classification is with the Split Window feature. Right-click a classification or a gene list created with k-means clustering and select **Split Window > Both**. The genome browser will divide into several smaller displays. (You can also choose vertically or horizontally.)

Copyright 1998-2001 Silicon Genetics

Self-Organizing Maps

The self-organizing map (SOM) is a clustering technique similar to k-means clustering, but SOMs, in addition to dividing genes into groups based on expression patterns, illustrate the relationship between groups by arranging them in a two-dimensional map. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another. SOMs were invented by Tuevo Kohonen (1991, 2000) and are used to analyze many kinds of data. Applications to gene expression analysis were described by Tamayo, et al (1999).

GeneSpring's self-organizing map algorithm begins by creating a two-dimensional grid of nodes in the space of gene expression. In each iteration, one gene is selected and all of the nodes within a user-defined "neighborhood" are moved closer to it. This process is repeated with each gene in the selected gene list until the maximum number of iterations has been reached. With each iteration, the "neighborhood radius" is incrementally reduced and nodes are moved by smaller and smaller amounts to produce convergence. In this way, the grid of nodes is stretched and wrapped to best represent the variability of the data, while still maintaining similarity between adjacent nodes. After the iteration is complete, genes are assigned to the nearest node, and a display grid of gene expression graphs is generated, corresponding to the initial grid of nodes.

To Create a Self-Organizing Map

1. Select **Tools > Clustering**. The Clustering window will appear. Under Clustering Method, select **Self-Organizing Map** from the drop-down menu.
2. Choose a gene list from the Gene List folder in the mini-navigator, right-click the list, and select **Set Gene List**. To remove a gene list, select the list in the *Genes to Use* box and click **Remove**.
   - To restrict the genes in the selected list, right-click an experiment or gene list in the navigator and select a restriction. For information on restrictions and how to apply them please refer to "Filter Genes Analysis Tools" on page 4-1.
   - To remove genes that may skew the clustering results due to missing measurements, click the **Discard Genes With No Data for Half The Conditions** box.
3. To add an experiment or condition, click on the experiment or condition in the Experiments folder in the mini-navigator, click the **Add** button and enter a weight in the New Experiment dialog box. The weight of a condition or experiment is a measure of the influence it has on the correlation distance, e.g. an experiment with a weight of 2.0 will be twice as influential as one with a weight of 1.0. To remove an experiment or condition, click on the experiment or condition under *Experiments to Use* and select **Remove**.
4. Choose the number of rows and columns in your grid. The default settings for the fields described in steps 5., 6., and 7. are based on the number of genes and conditions in your experiment. To return to the default settings after having changed these values, click the **Default Values** box at the bottom of the Clustering window. A good way to estimate the optimum number of rows and columns is to try to predict how many distinct classes of genes are affected by the conditions in your experiment. With small data sets, the algorithm may generate a number of empty nodes. To avoid this, you might try using a smaller grid.

Copyright 1998-2001 Silicon Genetics

5. Choose the number of iterations. This parameter controls how many times each gene is examined. If there are 10,000 genes and 60,000 iterations are specified, then each gene will be examined six times.
6. Choose the starting neighborhood radius. This parameter controls how many nodes move toward a data point at the beginning of the iteration, and therefore how similar the profiles will be for each node. As the iteration proceeds, the neighborhood radius decreases smoothly, so that points move more independently later in the process. The neighborhood radius is expressed in terms of Euclidean distance in grid units relative to the abstract grid of the expression patterns. (This is different from the distance between nodes in gene expression space.) For instance, point 1,2 is one unit away from 1,3. If you make the neighborhood radius very small (less than 1) each point will always move independently, and adjacent clusters will not be related. If you specify a very large neighborhood radius, initially all the nodes will move toward every data point, and the grid will act as if it is very "stiff", with more similarity between node results, but less flexibility to explore the variations in the data.
7. Click **`Start`**. When the analysis finishes, the Choose Classification Name window will appear.
8. Despite the name of the window, you can save the result either as a classification or as gene lists by selecting one of the two **`Save Classification as`**: radio buttons. Select a name for you classification/list folder and click **`Save`**.

Viewing SOMs

SOM results are best shown using the Split Window feature. Each graph contains the genes associated with a SOM node. Node numbers are shown in the upper right corner of each plot.

Copyright 1998-2001 Silicon Genetics

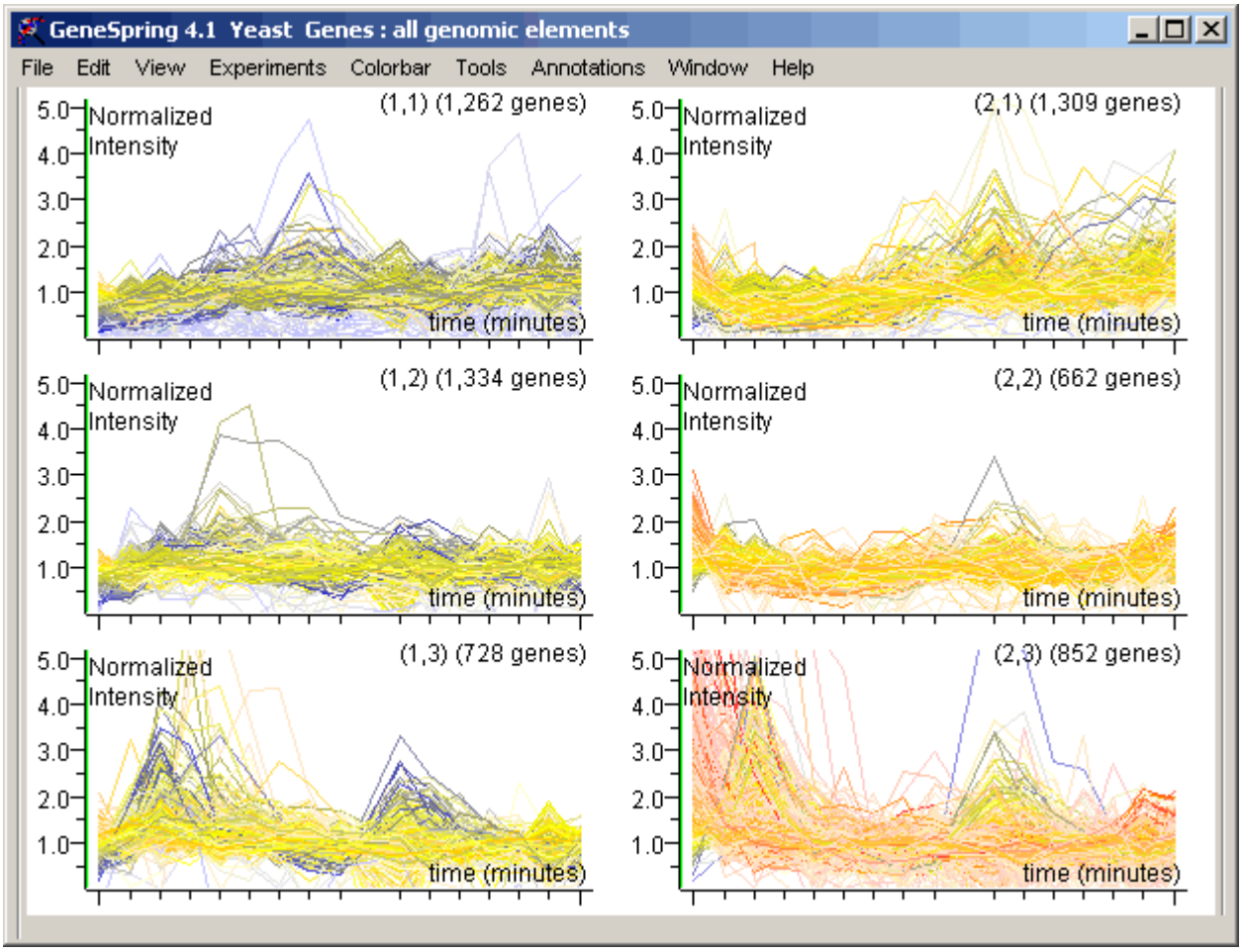


Figure 5-6 A 3x2 SOM of the "Yeast cell time series (no 90 min)" experiment

If you have selected many panels, you may want to hide the horizontal and vertical labels for easier viewing. Right-click the genome browser and select an option from the Options submenu. You can also increase your viewing space by selecting **`View > Visible > Hide All`**.

If you use a SOM to produce a classification, you can get details about the classification from the Classification Inspector. For information about the Classification Inspector, see "Classification Inspector" on page 3-46. To recreate your SOM graph, right-click the SOM classification or the folder of gene lists in the navigator and select **`Split Window > Both`**.

SOM References

Kohonen, T. (1990). The Self-Organizing Map. *Proc*. IEEE 78(9):1464-1480.

Kohonen, T. (2000). *Self-Organizing Maps* (Third Edition). Springer Verlag. Berlin.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps; Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* USA 96:2907-2912.

Copyright 1998-2001 Silicon Genetics

The Class Predictor

The Class Predictor is designed to predict the value, or "class", of an individual parameter in an uncharacterized sample or set of samples. It does this in two steps. First, the Class Predictor algorithm examines all genes in the training set individually and ranks them on their power to discriminate each class from all the others. Next it uses the most predictive genes to classify the "test set" (i.e. the set where the parameter value of interest is unknown). For example, you could attempt to diagnose the leukemia type of a leukemia patient with the Class Predictor by using expression data from patients whose leukemia type was known. You can also use the Class Predictor simply to find genes whose behavior is related to a given parameter by examining the list of predictor genes.

The list of predictor genes is assembled by ordering all the measurements for a given gene according to their normalized expression levels. For each class (parameter value), the predictor places a mark in the list where the relative abundance of the class on one side of the mark is the highest in comparison to the other side of the mark. The genes that are most accurately segregated by these markers are considered to be the most predictive. A list of the most predictive genes is made for each class and an equal number of genes are taken from each list.

To make a prediction, the class predictor uses the k-nearest-neighbor method. It selects "k" number of samples near (as measured in Euclidean distance) the unclassified sample, and for each class, computes a P-value that is the likelihood of finding the observed number of this class within the neighborhood members by chance given the proportion of the classes in the training set. The class with the lowest P-value is assigned to the unclassified sample.

You can specify a P-value cutoff, or threshold, such that if there is not sufficient evidence in favor of a particular class, no prediction will be made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.

To use the Class Predictor

1. Select **`Tools > Predict Parameter Values`**. The Predict Parameter Values window will appear.
2. Open the Experiments folder in the mini-navigator and click your training set (the set of samples for which the parameters are already known). Click the first **`Set`** button.
3. Click your test set (the set where the parameter value of interest is unknown), and click the second **`Set`** button.
4. Open the Gene Lists folder in the mini-navigator and click a gene list to be used in the selection process. Click the third **`Set`** button.
5. Specify a parameter type in the **`Parameter to predict`** box.
6. Choose a **`Maximum Number of Genes`** to be used in the prediction.
7. Specify a **`Number of Neighbors`**. Generally, this number should be no more than half the size of a single class, and no less than 10.

Copyright 1998-2001 Silicon Genetics

8. Specify a **P-value Cutoff**. The P-value cutoff is a threshold such that if there is not sufficient evidence in favor of a particular class, no prediction will be made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
9. Click **Predict Test Set** to make a prediction or **Crossvalidate Training Set** to evaluate how well the prediction rule can be used to predict the parameter values of the training set.
10. Selecting **Save Minimal Experiment** saves an experiment containing all of the samples in your training set, but including only the predictor genes. This is useful if you are making multiple predictions using the same training set and don't want to waste time recalculating the predictor list each time. The minimal experiment will be saved in your Experiments folder. The **Save Predictor Genes** button saves a list of your predictor genes. Genes are ordered according to their predictive values. The gene list will be saved in your Gene Lists folder.

Interpreting the Results of a Prediction

The Prediction Results window will appear after you have made a prediction or validated a training set. For convenience, not all of the prediction statistics are visible until you click the Show Details button at the bottom of the window.

- **True Value—**the true value of the class of each sample, as calculated when the parameter for the test set is already known. Compare this with the value in the Prediction column to validate your training set.
- **Prediction—**the predicted class.
- **P-value ratio—**the P-value ratio, or the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
- **Class counts—**the individual class counts for each sample.
- **P-value—**probability that individual class counts were found by chance.

The Class Predictor is designed for experiments with at least 20 or so samples in each class. It is possible to use the Predictor when you have very small sample sizes if you disable the P-value cutoff function. For sample sizes of less than 5, please specify 1 or 2 number of neighbors and specify 1 in the P-value cutoff field.

Copyright 1998-2001 Silicon Genetics

Chapter 6 Exporting GeneSpring Data

You can save a GeneSpring image and import it into a graphics or other program, where you can polish it and format it for publication. GeneSpring saves images of pathways, Venn diagrams, the genome browser, and the colorbar as *.pct* files, which can be imported into Microsoft® PowerPoint®, Word®, Publisher®, Excel®, CorelDRAW®, and Adobe® Illustrator® among other programs.

To Save a Genome Browser Image

1. Display the image you wish to save in the genome browser. This may be an image of a pathway.
2. Select **`File > Save Image`** and choose **`Browser`**. The Setup Graphic Size window will appear.
3. Choose an image size from the Overall size pull-down menu. You will have the following options:
   - **Original Image Size:** lets you save the image exactly as it appears in the genome browser.
   - **Original Aspect Ratio:** allows you to change the image size, but maintain the original width-to-height ratio displayed in the genome browser.
   - **US Letter:** 8.5 by 11 inches.
   - **US Legal:** 8.5 x 14 inches
   - **A4:** 8.3 x 11.7 inches
   - **3 Foot by 5 Foot Poster:** 3 ft. by 5 ft.
   - **Custom:** allows you to save to any size up to 450 inches by 450 inches.
4. Choose a Margin Size. If you choose Custom, you will need to enter a percentage in the Enter percentage box.
5. Choose a Mode - either landscape or portrait.
6. Click **`OK`**. A Save As window will appear. Choose a directory, type in a file name and click **`Save`**.

Note that you may need to save your file as a large custom size, such as 150x150 inches, to ensure all your data is included in the saved image. Note also that your image will be saved as a vector image, which is expandable, and that data that is too small to see in the genome browser will be saved in most cases, and will reappear when you expand the image. Be aware that images containing a very large number of genes can require an exceptional amount of memory. The fewer genes included in an image, the smaller the image file, and consequently the easier the image will be to open and manipulate in another program.

Copyright 1998-2001 Silicon Genetics

To save the Colorbar or Venn Diagram

1. Display the colorbar or Venn diagram you wish to save in the display window.
2. Select **File > Save Image** and choose Colorbar or Venn Diagram. A Save As window will appear.
3. Choose a directory and file name and click **Save**.

To save the Entire GeneSpring window

- **Windows PC—**Press the **Alt** and **Print Screen** keys simultaneously to copy a picture of the current active window. Paste the image into any program that accepts graphics and save it.
- **Macintosh—**Press ⌘-**Shift-4-Caps Lock** simultaneously. The cursor will change to a bull's-eye. Click on a GeneSpring window to save the image as a file on your hard drive called "Picture". You will need to rename this file.

To save the Entire Computer Screen

- **Windows PC—**Press the **Print Screen** key to save an image of your entire computer screen. Paste the image into any program that accepts graphics and save it.
- **Macintosh—**Press ⌘**-Shift-3** simultaneously to save an image of your entire computer screen. The image will be saved as a file on your hard drive called "Picture".

Saving Pictures and Printing

You can print an image of the genome browser, the genome browser with the colorbar, or the display window. Such images can be useful for reports or handouts. Please use a high-resolution color printer to print GeneSpring images.

To Print an Image of the Genome Browser and/or Colorbar

1. Select the **File > Print Image** command.
2. Choose from the following options:
   - **Browser:** prints only the genome browser
   - **Browser and Colorbar:** prints the genome browser and colorbar
   - **Colorbar:** prints only the colorbar
3. Select a printer and click **OK**.

Copyright 1998-2001 Silicon Genetics

To Print an Image of the Display Window

For Windows PC:

1. Hold the **Alt** and **Print Screen** keys down simultaneously. This will copy a picture of the active window only.
2. Paste into any program that accepts graphics.
3. Print.

For a Macintosh:

1. Hold the **Command**-**Shift-4-Caps Lock** keys down simultaneously. The cursor will change to a bull's-eye.
2. Release the keys and use the mouse to click on the window. This will create a screenshot of your window (you will hear the sound of a snapshot). The screenshot will be saved on your hard drive with the name "Picture".
3. Open the picture and print.

Exporting Gene Lists out of GeneSpring

You can make gene lists and annotated gene lists available to another application. An annotated list includes functional descriptions, as well as standard deviation, standard error and other information associated with the gene list.

To copy a gene list

1. Select the gene list you wish to copy from the Gene Lists folder in the navigator.
2. Select **Edit > Copy > Copy Gene List**.
3. Paste the list into another application, such as a spreadsheet program.

   Or,

1. Open the Gene List Inspector. (Double-click a gene list or right-click and select **Inspect**.)
2. Click the **Copy to Clipboard** button.
3. Paste the list into a new application.

Both of these methods will export the default interpretation of your gene list.

To copy an annotated gene list

1. Select the gene list in the Gene List folder in the navigator.
2. Select **Edit > Copy > Copy Annotated Gene List**. A menu will appear.
3. Choose an experiment interpretation from the **Copy based on interpretation** pull-down menu. (See "Changing the Experiment Interpretation" on page 2-17 for information on experiment interpretations.)

Copyright 1998-2001 Silicon Genetics

4. Choose options on the Copy Annotated Gene List window by checking or unchecking the boxes.
5. Click the **Copy to Clipboard** button.
6. Paste the list into another application.

To save an annotated gene list

1. Select a gene list from the Gene List folder in the navigator.
2. Select **Edit > Copy > Copy Annotated Gene List**. A menu will appear.
3. Choose the experiment interpretation from the **Copy based on interpretation** pull-down menu. (See "Changing the Experiment Interpretation" on page 2-17 for information on experiment interpretations.)
4. Click the **Save to Disk** button.
5. Choose a name and location to save your gene list.

The resulting text file can be opened in any program that accepts tab deliminated text, such as spreadsheet and word processing programs.

Annotation Options

Your options for copying and saving information with an annotated gene list are listed in the Copy Annotated Gene List window. Descriptions of these items can be found by clicking the **Help** button. The type and amount of information listed will vary depending on your genome and the way that genome was loaded into GeneSpring.

- **Gene List Associated Value—**The values (if any) that GeneSpring has associated with this gene list. This column will only show up if you have associated values. Refer to "Adding an Associated Number Restriction" on page 4-9 for more details on the types of numbers GeneSpring attaches to gene lists.
- **Gene List Note—**Any notes attached to a gene list. This options appears only if a gene list note exists.
- **Systematic Name—***The systematic name is not listed in the Copy Annotated Gene List window, but is automatically saved in the first column of a gene list.* It appears when you paste or open the gene list in a new application.

Identifiers

- **Common Name—**A non-systematic way of referring to a gene.
- **Synonyms—**Other names entered for your gene list.
- **GenBank—**A gene's GenBank Accession Number, if known.
- **EC—**A gene's EC (Enzyme Commission) number, if known.
- **PubMed—**A gene's PubMed identifier.
- **DB id—**A reference used to identify a gene within GeNet.

Copyright 1998-2001 Silicon Genetics

Normalized Data

- **Average—**The mean of any normalized replicates in the experiment.
- **Minimum—**The minimum normalized signal values for each gene.
- **Maximum—**The maximum normalized signal values for each gene.
- **Flags—**Any measurement flags associated with genes in the list.
- **Standard Error—**The standard error of the normalized values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the normalized values for each gene.
- **t-test p-value—**The t-test p-value which measure the significance of differential gene expression in each condition.

Logarithm or Fold Change

- **Average—**The mean of any normalized replicates in the experiment.
- **Minimum—**The minimum normalized signal values for each gene.
- **Maximum—**The maximum normalized signal values for each gene.
- **Standard Error—**The standard error of the normalized values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the normalized values for each gene.

Raw Data

- **Average—**The mean of any raw data replicates in the experiment.
- **Minimum—**The minimum raw data signal values for each gene.
- **Maximum—**The maximum raw data signal values for each gene.
- **Standard Error—**The standard error of the raw data values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the raw data values for each gene.

Control Value

- **Average—**The mean of any control value replicates in the experiment.
- **Minimum—**The minimum control value signal values for each gene.
- **Maximum—**The maximum control value signal values for each gene.
- **Standard Error—**The standard error of the control values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the control values for each gene.

Copyright 1998-2001 Silicon Genetics

Annotations

- **Description—**A gene's description, if known.
- **Phenotype—**A description of a gene's phenotype, if known.
- **Function—**A description of the function of a gene's product, if known.
- **Product—**The protein product coded for by a gene, if known.
- **Map Position—**A gene's mapping information.
- **Chromosome—**The chromosome on which a gene is located, if known.
- **Keywords—**Keywords associated with a gene, if known.
- **Custom Field 1, Custom Field 2, Custom Field 3—**Whatever information you may have placed here for your own use.

Publish to GeNet

GeNet™ is a web database designed to distribute and visualize any organisms' gene expression data from microarrays and related technologies. It allows researchers to publish raw text data, images, annotations, and the results of analyses in any file format.

For details about GeNet, its installation and troubleshooting, please refer the *GeNet User's Guide*. You must have several different pieces of software to make GeNet work, so please consult with your system administrator as needed.

Upload to GeNet

Start GeneSpring as usual. Position your cursor over a data object in the navigator you would like to upload and right-click. Select **`Publish to GeNet`** from the pop-up menu.

You can publish all of the data objects present in GeneSpring to GeNet.

GeNet can generate magnifiable and selectable images including:

- bar graphs plot
- classification
- graph by gene
- line graphs
- ordered lists
- pathways
- physical position graphs (where available)
- scatter plots
- trees

  All of these types of data will be referred to as *data objects*.

Copyright 1998-2001 Silicon Genetics

GeNet can also generate reports including:

- experiment reports
- gene list reports
- annotated data

Every folder, genome, list, tree etc. that can be uploaded to GeNet will have a **Publish to GeNet** menu item in its right-click pop-up menu.

Once selected, the GeNet Upload window will appear.

Type in any necessary information. Once you click the **Upload** button you will see a new dialog box. This box will contain information on the progress of the upload. Each item (if you are uploading an entire folder) will have its own line.

If GeNet is not available or if you are unable to load data for another reason, you will get an error message. If you specify a nonexistent destination directory, GeNet will create one.

If you are having trouble uploading, ask your administrator to check and make sure your default directory exists. It can easily be added if it does not exist. Depending on the initial set up of GeNet, you may not have access to every directory.

Once your upload is complete the upload status box will say it is complete. Click the **Close** button or the small **x** in the upper right corner.

Uploading Genomes to GeNet

You must have administrator access privileges to upload genomes to GeNet. If you cannot upload genomes and feel you should, please contact your system administrator. To upload a genome to GeNet, go to **File > Publish Genome to GeNet**. Type your identification into the screen as necessary and click the **Upload** button.

When uploading genomes to GeNet, there is an **Update Existing Genome** checkbox under your password. This field is always unselected by default. Normally, if you try to upload a genome which is already present on the server, it simply gives an error message. If you select this option by clicking in the box, GeNet will update the genome to make it like the genome you are uploading. Specifically, GeNet will:

- add new genes to the genome
- change annotations on existing genes
- change the lists of hypertext links for genes and experiments

However, GeNet will *not* remove genes from the genome, since there might be gene lists, experiments, etc. which involve those genes.

Copyright 1998-2001 Silicon Genetics

Using GeNet

To view your data, or someone else's, on GeNet you will need to start your usual web browser and go to the web page specified by your administrator. Enter your user GeNet ID and password to log on. Select a genome to view and click **Continue**.

Loading Data from GeNet

You can download data objects from GeNet and manipulate them on your local copy of GeneSpring.

1. From the main GeneSpring window, select **File > Load Data from GeNet**. You will be prompted for your GeNet user name and password.
2. Type in your GeNet user name and password. Click **OK**. A window may appear informing you GeneSpring is catching data. Click **OK** or wait.

In a moment, GeneSpring will have passed all the data it needs and you will have several new folders in the navigator. Each top level folder (Gene Lists, Experiments, Gene Trees and so on) will contain a new folder called *GeNet* containing the data just collected from GeNet.

The folders created in this feature are "links" to GeNet. The data in GeNet is not really downloaded to your local hard drive, as that would take up too much space.

If you use the **Load Data from GeNet** command twice in the same session, you may get the folder duplicated within GeneSpring. To avoid this, please shut down GeneSpring between uses.

All items being viewed from GeNet appear in an *italic font* within the navigator.

You cannot delete a GeNet data object from the server, but you can remove it from your navigator by right-clicking over the data object and selecting **Delete List** or similar command from the pop-up menu.

Copyright 1998-2001 Silicon Genetics

Appendix A Help

Contacting Silicon Genetics' Technical Support

You may contact Silicon Genetics' Technical Services Department at 650-367-9600 or support@sigenetics.com.

There is a great deal of current, useful information on the Silicon Genetics' website, select **`Help > Frequently Asked Questions`** to launch your browser and reach http://www.sigenetics.com/GeneSpring/faq/index.html

The Help Menu

The Help Menu is located on the right of the menu bar.

GeneSpring Basics Instructional Manual

You can download this file from the web and print it (if you wish) as a PDF document. The tutorial covers many basic topics of GeneSpring.

Manual

Selecting the Manual will launch your browser and take you to C:\Program Files\SiliconGenetics\GeneSpring\docs\GeneSpringMainScreen.html. The *GeneSpring User Manual* is a PDF document you can save or print.

FAQ

Selecting the Frequently Asked Questions will launch your browser and take you to http://www.sigenetics.com/GeneSpring/faq/index.html

Version Notes

Selecting this will launch your browser and takes you to C:\Program Files\SiliconGenetics\GeneSpring\docs\VersionNotes.html. This page should have all the version notes for your version of GeneSpring.

Copyright 1998-2001 Silicon Genetics

Update GeneSpring

Selecting **`Update GeneSpring`** will bring up a window where you can agree to the conditions and get a new version of GeneSpring if your license is still active.

You can also automatically update the manuals that accompany GeneSpring. The manuals are typically published at HTML or PDF documents and it is recommended to update them every time you update GeneSpring.

Selecting this item will launch your browser and take you to a webpage to download a new copy of GeneSpring. Make sure it is saved in the correct folder.

Silicon Genetics on the Web

Selecting this will launch your browser and take you to http://www.sigenetics.com/GeneSpring/index.html. There should be manuals and information on workshops designed to help you use GeneSpring more effectively.

GeNet Database

Selecting this item will launch your browser and take you to a webpage describing GeNet™. You can download a demo copy of GeNet™ from that page. You will also see other commands to upload or download with GeNet. Please see “Publish to GeNet” on page 6-6 or the *GeNet User Manual*.

Register for a Workshop

Selecting this will launch your browser and take you to Silicon Genetics training page. Here you can take advantage of Silicon Genetic’s many training options.

System Monitor

This item will bring up the Java System monitor with information about free memory and what is currently happening on your computer. If you are running low on memory, GeneSpring will bring up a warning box.

About

Selecting **`Help > About`** will bring up the initial graphic of GeneSpring, showing you the version number, demo expiration date and other useful information.

Also, only for Macintosh users there is a confirmation dialog appearing at the closing of the last browser window.

Copyright 1998-2001 Silicon Genetics

Appendix B Preferences Window

The preferences screen allows you to change GeneSpring's global preferences. Note that some changes may not take effect in the currently open window in the current run. All of these preferences will take effect when GeneSpring is restarted.

Select **`Edit > Preferences`**. To change any options in the Preferences window, select the drop-down menu and choose the appropriate item.

Data Files

Here you can set the defaults of what you would like to see when GeneSpring opens. By setting the defaults in this box, you can have GeneSpring open directly to your chosen experiment.

- **Data Directory:** The default directory genome that opens at startup. Use the browse button to select the settings.
- **Default Genome:** To change the default genome that is loaded when GeneSpring first starts, enter the name of a genome in this field.

Database

If you plan to store your experiment's expression data in a database, the Database panel allows you to specify the method GeneSpring will use to extract data from an ODBC compliant database. The drop-down menu (selecting the black arrow will produce another option, **`Parameters appearing to be numeric list individually`**) allows you to specify how GeneSpring will assign the parameters for a series of numeric values in your database. In addition, you will need to specify the fully qualified classname of the driver in the JDBC driver field.

Copyright 1998-2001

Color

The Color panel allows you to change the colors GeneSpring uses to represent different types of data and other screen elements. In this box you may change the color defaults to any of the listed colors until you find a combination you like and is easy for you to see on the screen.

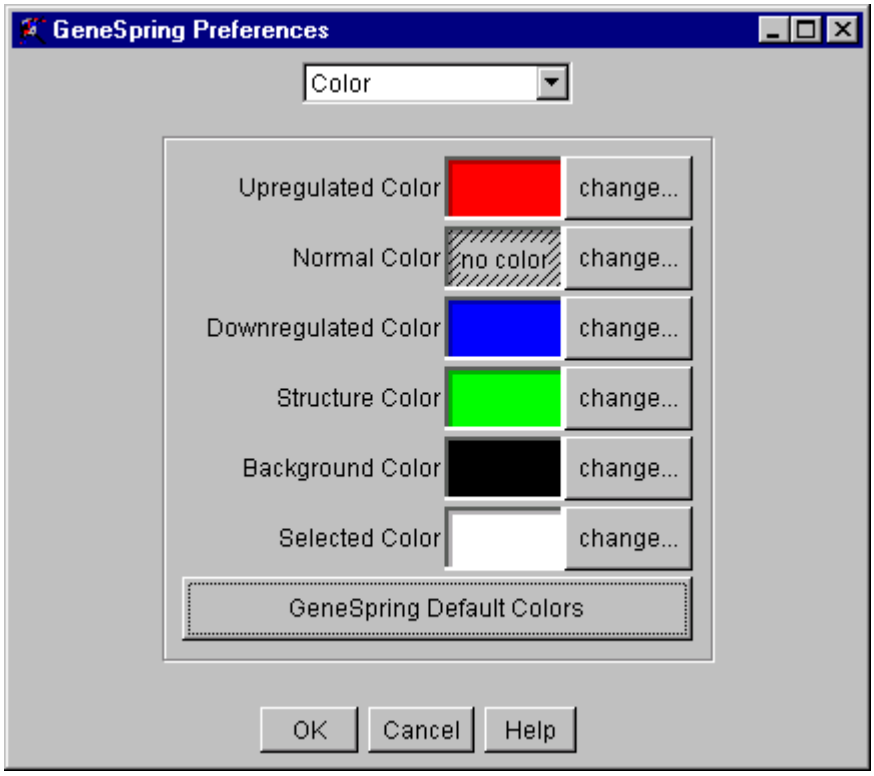


Figure 4-1 The Colors section of the Preferences window

- **Upregulated Color:** The Upregulated Color is the color that will be used to display genes greater than or equal to the *High Expression* value selected for the current color bar. The default for this color is red. The brightness of the color depends on the trust associated with it. Please refer to “Trust” on page 3-32.
- **Normal Color:** The Normal Color is the color used to represent genes having a normalized expression value of one. The default for this color is yellow.
- **Downregulated Color:** The Downregulated Color is used to display genes less than or equal to the *Low Expression* value selected for the color bar. The default for this color is blue.

Over- and under-expression color refers to the coloring of genes as shown in the genome browser and color bar. You can change the definitions of overexpressed (upregulated) and underexpressed (downregulated) genes by right-clicking over the colorbar in the main genome browser and resetting the defaults. Please refer to “Changing the Experimental Data Range” on page 3-36 for more details on this topic.

Copyright 1998-2001 Silicon Genetics

- **Structure color:** The Structure Color is used for the ConditionLine and for the lines between the genes in the Physical Position View, the Tree lines, the Ordered List lines, etc.
- **Background Color:** The Background Color defines the color behind the genes and other elements in the genome browser.
- **Selected Color:** The Selected Color is used for selected genes, gene names, and axes. For this, you will probably want the greatest contrast with the background color.

For more information on the various color options on GeneSpring, please refer to “Changing the Coloring Scheme” on page 3-31.

Specific Color Definition

A new feature in GeneSpring version 4.1 is the ability to define *exactly* what color you would like to use in the genome browser. If your printer requires exact color definitions, your life should be much easier after this.

To change or adjust a color in GeneSpring, select the **Change** button next to its element in the Preferences Colors window.

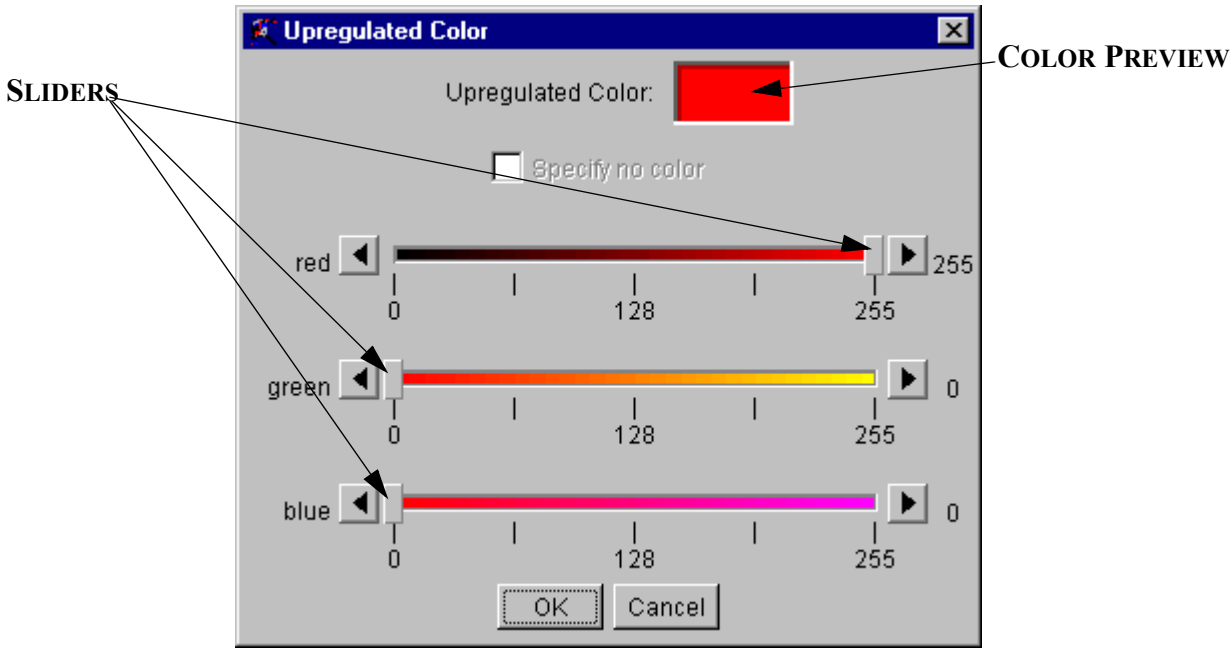


Figure 4-2 Color creation in the Preferences window

Using your cursor, click over any slider and move horizontally to adjust the color. Keep an eye on the color preview box and stop moving the cursor when the desired color is reached. Click **OK** to accept the new color.

Copyright 1998-2001 Silicon Genetics

Gene Labels

This function allows you to specify how you would like to name your genes in the genome browser. The defaults are systematic name and common name.

This feature is particularly useful in the Scatter plot.

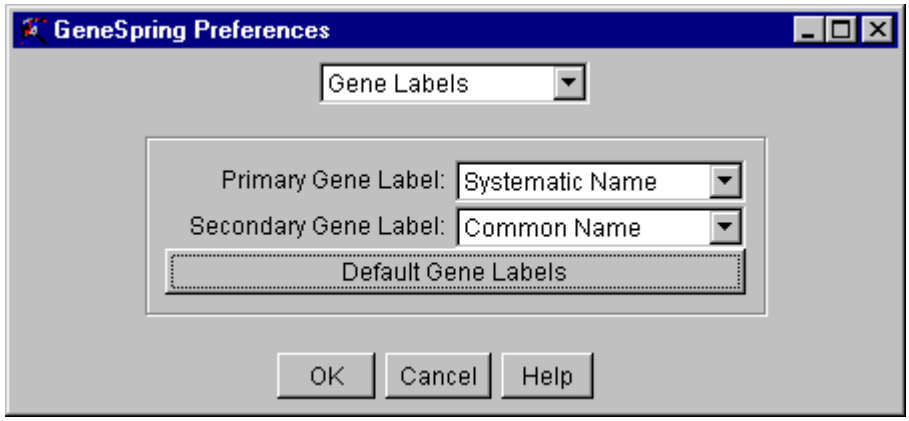


Figure 4-3 Gene Labels details in the Preferences window

Browser Details

In this box you can set the defaults for your web browser in case you want to use a particular browser for the GeneSpring applications. You will only need the use the Browser assignment field if you are using an obscure web browser that requires and argument.

The Firewall Details box

If your company has a firewall to prevent unauthorized use of the internet, you will need to use this box to get through it. You may need to contact your System Administrator for details about your firewall.

Copyright 1998-2001

The System Preferences

The System panel allows you to specify a number of different parameters about networking and memory usage.

- The License Manager field allows you to specify the IP address of the machine that dispenses concurrent licenses.
- The GeNet Address field contains the URL of GeNet in your company or institution.
- The Desired Memory field sets the amount of RAM GeneSpring will attempt to use. If this field is set too high (with respect to the total available memory), unnecessary disk caching will occur and performance will be slowed.
- The Disk Cache Size field specifies the amount of hard disk space GeneSpring uses to store HTML pages accessed by the GeneSpider or by other internet-based search functions.

The Miscellaneous

The Miscellaneous panel contains a grab-bag of defaults to customize your GeneSpring installation.

- The ***Default Correlation*** field specifies the default minimum correlation coefficient that appears near the Find Similar button in the Gene Inspector window.
- The ***Restrict Gene List Searches*** drop-down menu allows you to limit the lists GeneSpring examines when searching for similar lists in the Gene Inspector window and during Tree building.
- The ***Default Font*** field allows you to specify the name, style, and point-size (in this order separated by hyphens) for most of the text within the GeneSpring window. When you first install GeneSpring, the name and style fields are left blank, and only the point-size is specified (*e.g.* `--9`). An example of an alternative font specification might be, `Serif-Bold-12`. The available font styles are “plain”, “italic”, “bold”, and “bolditalic”. The available font names differ depending on what JVM you are using. Start with the generic font classes, “Serif”, “SansSerif”, “Monospaced”, and “Dialog”. Please be aware, some virtual machines support the use of explicit names for fonts that are available to the operating system.
- The ***Unique ID prefix*** field allows users to specify an alphanumeric prefix that will be appended to the identifier field within data files. If you commonly share genelist files between different GeneSpring installations, it is a good idea to give each installation different ID prefix so GeneSpring is not confused by genelists with similar identifiers.
- The ***Your Name***, ***Your Group Name***, and ***Your Email*** fields contain the text that is contained in the HTML files that go into your data directories.

Copyright 1998-2001

Copyright 1998-2001

Appendix C Genome Wizard

Each and every genome known to GeneSpring must have its own *.genomedef* file. You can create a *.genomedef* file by hand (please refer to "The .genomedef File" on page I-1), by using the Autoloader (please refer to "Creating a Genome through the Autoloader" on page 2-7) or by using the Genome Wizard.

The Genome Wizard will guide you through the steps of creating a *.genomedef* file. Most of these panels are fairly self-explanatory. Most Wizard panels will take up most of your screen. This is to prevent any necessary boxes from being shrunk to a non-visible size. You can change the size of any panel in the usual manner of grabbing an edge with the cursor and dragging, but it is recommended you leave them at the large size. You may not see every panel discussed here as you go through the Genome Wizard as the Genome Wizard will modify itself depending on your answers.

1. Select **`File > New Genome Installation Wizard`**. The New Genome Installation Wizard panel will appear. In this window you need to tell GeneSpring the name of the genome you are installing. To name a genome:

   a. Place the cursor in the Organism Name box.

   b. Type the name of the organism as you wish it to appear in GeneSpring. This name can be anything, but a sensible, memorable name is recommended. GeneSpring will remember this name with the capitalization and the spelling you use here.

   c. Click the **`Next`** button to move forward to the next panel.

2. Genome Data Directory panel will appear. In this panel you can select or create a new directory. GeneSpring will bring up a default directory, named the same as the organism you just entered. If you type in the name of a non-existent directory GeneSpring will create it for you. Later you can use the Wizard to select various files and GeneSpring will copy them into this directory automatically. See "Raw Data" on page K-1 for the correct format of the raw data files. To enter the directory:

   a. Type the complete directory pathway name in the *Specify directory* box.

   If you already have a directory for the organism you named in previously, GeneSpring will ask you to define a subdirectory. If you are starting a new species directory this will be unnecessary.

   Or, if you have already created a directory as specified in "Creating Folders for New Genomes" on page H-1, you will need to type in or browse to find that directory. To browser to a directory:

   a. Click the **`Browse`** button. A dialog box will come up showing the *data folder* in GeneSpring. Before you begin browsing, look at the folder to make sure you are in the folder you want.

   b. Find the file directory (folder) containing your raw data files.

Copyright 1998-2001 Silicon Genetics

c. Click the directory file (folder). This opens the directory. You should see your raw data files within this directory.

d. Click the **Save** button. This writes the pathway in the *Specify directory* box of the Genome Wizard.

When you click the **Save** button in the *Browse directory* window, the *File Name* box in the window contains the file name "[Dummy Name, leave alone]". This is what the window is supposed to look like when you click the **Save** button. If you accidentally click one of the files within the genome's directory, the name in the File name box changes. Then, when you click the **Save** button you will get an error message.

Click the **Yes** button of this error message; this does not replace the raw data file, it simply enters the directory of the correct file into the *Specify directory* box of the Genome Wizard.

Click the **Next** button in the Genome Wizard to move to the next panel. If you click **Next** without specifying your genome directory, then GeneSpring will create a directory for you in the GeneSpring\data directory. Directories automatically created in this way are named using the name of your genome. GeneSpring will automatically copy your files into this directory. You can select **File > New Window** to see the new files.

3. The Overall Genome Properties panel will appear. In this window you tell GeneSpring whether the genome you are entering has been sequenced, and if it has a circular genome.

   a. In the first box, select the **Yes** circle if your organism has been sequenced, otherwise leave the **No** circle selected.

   b. In the second box, select the **Yes** circle if your organism is a circular genome, like bacteria, plasmids, and viruses. If it is, GeneSpring will display it as a circle in the physical position display. Leave the default setting of **No** selected if your organism does not have a circular genome.

   c. Click the **Next** button to move forward to the next panel.

4. The GenBank Data File panel will appear. While GenBank offers several different files for their complete genomes, GeneSpring can *only* read their *.gbk* files. In this panel you tell GeneSpring if you are using a GenBank file as your data source, and if so, what the file is named. An EMBL file may be used in place of a GenBank file. For the purposes of this panel, treat the EMBL file as if it were a GenBank file; answer **Yes** to having a GenBank file and enter the file name and pathway of the EMBL file where it asks for the GenBank file name. You may need to download a GenBank file, please see "GenBank or EMBL Files" on page H-4. To indicate you have a GenBank or EMBL file:

   a. Select **Yes**. If you are not using a GenBank or EMBL file, leave the **No** circle selected and go on to the next panel.

   b. Either type the complete file name and pathway of your GenBank/EMBL file in the *Enter filename* box, or click the **Browse** button. This brings up the browser window.

   c. Look at the folder listing to make sure you are in the folder you want.

   d. Click the GenBank or EMBL file for this organism.

Copyright 1998-2001 Silicon Genetics

e. Click the **Open** button. This enters the complete pathway and file name of the selected file in the *Enter filename* box of the Genome Wizard.

Once you indicate you have a GenBank/EMBL file, then this panel will not let you move forward until you have entered the file name of your GenBank/EMBL file in the *Enter filename* box. When you use the **Browse** button to select the GenBank/EMBL file, click once in the Wizard panel to make it the active window. Then click the **Next** button to go on to the next panel. If you do not use the browse feature, be very careful of spelling and capitalization errors, as GeneSpring attempts to locate the file before it allows you to progress to the next panel.

5. The Master Gene Table panel will appear. You will not see this panel if you are using a GenBank or EMBL file for your organism. Your Master Gene Table must be in a name list, name function, SGD or mapped format. Please see "What Format do these Data Need to be in?" on page H-1 for an example. This panel tells GeneSpring what the name of your Master Gene Table is, and what format it is in. The Master Gene Table is referred to as a "Gene List" file in this panel, because the list of gene names are the most important information contained in the Master Gene Table. To enter the Master Gene Table's file name, either type the complete pathway and file name of the Master Gene Table file, or:

   a. Click the **Browse** button. A window will appear. Look at the folder listed to make sure you are in the folder you want.

   b. In this new window, select your Master Gene Table file (for example, *ORF_table.txt*).

   c. Click the **Open** button. This enters the filename and pathway within the *Enter GeneList Filename* box of the Genome Wizard.

   The Master Gene Table file will be copied into the correct folder by GeneSpring. You will not be able to go to the next panel until a Master Gene Table file has been indicated. GeneSpring checks to make sure the file name you typed actually exists. Beware of spelling and capitalization errors because if GeneSpring cannot locate the file you indicate you will not be permitted to progress to the next panel.

6. The Genome Sequence File panel will appear. You will not see this panel unless you indicate in the *Overall Genome Properties* panel that your genome has been sequenced, and you are not using a GenBank or EMBL file. This panel tells GeneSpring where to find the sequence data. To do this, click the **Enter Genome Sequence File Name** box and type the complete file name and pathway or:

   a. Click the **Browse** button. A window will appear. Look at the listed folder to make sure you are in the folder you want.

   b. Select the *.seq* file containing your organism's sequence.

   c. Click the **Open** button. This enters the file name and pathway into the **Enter Genome Sequence File Name** box of the Genome Wizard.

   You cannot go onto the next panel until you have entered a file name. The sequence data file will be copied by GeneSpring to the correct directory. The file you indicate in the Enter Genome Sequence File Name box must exist, or the Genome Wizard will not let you continue.

Copyright 1998-2001 Silicon Genetics

Beware of spelling and capitalization errors as GeneSpring needs to locate the file before allowing you to progress to the next panel.

7. The Additional Genetic Elements panel will appear. This table tells GeneSpring if you have a second table of genes. Generally a second table of genes is used if you want to add genetic elements to a GenBank or EMBL-defined organism. In this case the supplementary table of genes probably contains alleles, centromeres, or genes from strains differing slightly from the sequenced strain. To tell GeneSpring where to find the additional elements:

   a. Click the **Yes** circle to select it. If you do not have a separate table of genes file leave the **No** circle selected and go to the next panel.

   b. Either click in the *Enter Filename* box, and type the complete file name and pathway, or click the **Browse** button to select a file. Look at the listed folder to make sure you are in the correct directory.

   c. Click the table of genes file containing the extra genomic information.

   d. Click the **Open** button. This will insert the file information into the Enter Filename box.

   e. Click the arrow to the right of the **Select a file format** box. A menu will appear.

   f. Click the format used in the supplementary table of genes file. For a description of the four format options, see section "What Format do these Data Need to be in?" on page H-1.

   Once you indicate you have a file containing extra genomic elements, you cannot proceed to the next panel until you have indicated a file and a file format. Beware of spelling and capitalization errors when indicating the file name and pathway, as GeneSpring checks to make sure the file you name exists before letting you go on to the next panel.

8. The Links to Web DataBases panel will appear. This panel allows you to link GeneSpring directly to web-based data sources on your genes. You can create a link to a URL containing the name of one of your genes. If you would like to have any such links, select the **Yes** circle. In the **Enter number of links** box type the number of web databases you want to link the genes in this genome to. When you enter a number in this box, the number of "Button" lines in the table below changes. In the first column of this lower table (titled **Button label**) enter the name of the web database as you wish it to appear on a button within GeneSpring. In the right-hand column (titled **URL**), enter the URL of the database, with the systematic name of the gene replaced by a semicolon. If the semicolon representing the place the systematic name of the gene should go is at the end of the URL, it may be omitted. You can also have links using names other than the systematic gene name. To use one of these, attach a special character before the link name (in the **Button label** column). Do not put a space or other character between the special character and the link name. To use the common name, use a dollar character ($). To use the GenBank Accession Number, use a percent sign (%). To use the systematic name, less anything after a dash, use the dash (-).

   a. Select the **Yes** circle and the **Next** button if you have databases on the World Wide Web you would like to easily access from GeneSpring.

   If you want to place more buttons, you can change the number in the **Enter number of links** option. Then use the tab key to move through the Button Label table.

Copyright 1998-2001 Silicon Genetics

When you right-click the table in this panel of the Genome Wizard, there is no pop-up menu allowing you to cut and paste. You can still cut and paste URLs into the matrix fields by using the keyboard commands (for Windows® this is Ctrl+C and Ctrl+V). Cutting and pasting has a much higher success ratio as URLs are both spelling and case sensitive. GeneSpring will attempt to locate each URL you insert before it allows you to proceed to the next panel. This may be a problem if you are not connected to the internet when you are creating this genome. If this is the case you will have to skip this panel and add the web-links to the *.genomedef* file later. To add hyperlinks from GeneSpring, please see "Searching Internet Databases" on page 3-40.

For NT and Mac users, you should set the path to your usual browser, because GeneSpring can not automatically locate the default web browser on NT or Mac machines, which may cause you trouble in this panel. To set the path to the browser:

a. Select **Edit > Preferences**.

b. Select **Browser** from the drop-down menu.

c. In the *Browser path* box, either type the complete file name and pathway of the *.exe* file for your default browser, or click the **Browse** button to the right of the *Browser path* box. If you do this, a window will appear.
The default from the Preferences box may take you into the wrong folder. You will need to look for your default browser's files in your system directory. In a Windows NT environment your path may look something like this:
`C:\Program Files\Plus!\Microsoft Internet\IEXPLORE.EXE`

- Find and select the *.exe* file associated with your internet browser.
- Click the **Open** button in the *Browse* window. This writes the complete *.exe* file name and pathway in to the *Browser path* box of the Preferences window.

d. Click **OK** to close the Preferences window. The path to your browser should be set.

9. The Miscellaneous Settings panel will appear. This panel lets you alter the way the gene names are displayed.

   a. If you wish to force all of the systematic gene names to upper or lower case letters select the appropriate check box.

   It is perfectly acceptable not to select any of the check box options.

   b. Select **Next** to proceed to the next panel.

10. The Finished panel will appear. When you click the **Finish** button all of the answers you gave in the previous Genome Wizard panels are saved in a *.genomedef* file.

Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Appendix D The Experiment Wizard

Before you begin installing your new experiment you need to go through the Genome Installation Wizard to specify a new genome, if the genome for your experiment is not yet in GeneSpring so GeneSpring will correctly interpret what you are telling it. If you are not cutting and pasting data, you will need to create a folder called Experiments and place your experimental data files in that folder so they will be easy to find when you need them later in this process.

Files You will Need to Use the Experiment Wizard

An experimental data file is the main file needed for loading an experiment. Gene names need to be listed in the first column, one name per line, with the experimental data reported in subsequent columns. Viewed in a spreadsheet, it might look like this:

Gene Name	Control Strength in Experiment 1	Control Channel Strength	Background Signal	Background Signal for the Reference	Experiment Flag	Region
CLN1	510	110	10	10	P	A
MEP2	9	19	9	9	M	C

If created in a spreadsheet program, the file should be saved as a tab-delineated text file.

If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use commas and periods interchangeably.

GeneSpring can also read experimental data from databases via an ODBC link. Please refer to “Installing from a Database” on page E-1.

- **Pictures of the conditions during the experiment:** Pictures of a condition can be useful reminders of what was happening in an experiment at a given point in time. In GeneSpring, you can associate a maximum of one picture with each condition. Even with only a few pictures, GeneSpring will display the picture closest to the condition you are viewing. These pictures should be either *.gif* or *.jpeg* files.
- **Pictures of the Microarray plates:** At most there can be one array picture associated with each sample.

  These pictures should be either *.gif* or *.jpeg* files.

Copyright 1998-2001 Silicon Genetics

- **The Positive and Negative Control Files:** A positive control file and a negative control file are formatted in exactly the same way; their contents are different. Each file lists the control genes' names, one name per line:

```
Control Gene Name 1
Control Gene Name 2
Control Gene Name 3
Control Gene Name 4
Control Gene Name 5
Control Gene Name 6
. . .
```

This list of gene names is all either file should contain. There should not be any headlines or anything else in the file, only the gene names.

Briefly, you have negative controls in your experiment when there is DNA from a different genome than the one you are investigating on the array. You are using positive controls when there is DNA from a different genome than the one you are investigating on your array, and you add a known quantity of that different DNA to your sample. For a description of the possible normalizations to be done with these controls see “Normalizing Options” on page G-1.

The names of the positive and negative controls do not need to be listed in your Master Table of Genes. If they are listed, those genes will be colored gray (not measured) in the genome browser because they are used in normalization not measurement.

Once all your files are together, you can start the Experiment Wizard.

Copyright 1998-2001 Silicon Genetics

The Experiment Import Wizard

Most of the panels in the Experiment Import Wizard are fairly self-explanatory. This section is mainly designed to show the different possible appearances a panel can have, and add any notes about characteristics that are not obvious. The *Experiment Import Wizard* saves your experiment information as an HTML file. When you are entering a new experiment make sure the genome browser in the main GeneSpring window is displaying the genome the experiment refers to.

To initiate the Wizard, select **`File > Manual Load Experiment > Experiment Import Wizard`**. (If you are about to load an experiment very similar to an experiment you already have in GeneSpring, you can use the **`Experiment Import Wizard (like this experiment)`** to expedite the loading process. In this case "similar to" means the same genome, same file layout and similar conditions.)

1. The Welcome panel of the GeneSpring Experiment Entry Wizard will appear. This panel will contain some instruction on how to prepare for using the wizard, including the types of files necessary.

   Clicking the **`Help Pasting Data`** button will take you to a web page with information on pasting experiments directly into GeneSpring. Pasting is very easy (if your file is set up correctly) but it is not very flexible. Please refer to "Copying and Pasting Experiments" on page F-1 for more information. The Experiment Wizard is very flexible, and correspondingly more complex.

   The Welcome panel includes lists to remind you to create or gather your raw data files. There are five possible raw data files listed below; only the first one is necessary for loading an experiment. They should all be placed within the "Experiment" sub-folder of the relevant organisms described in "Where do I put my data?" on page K-8.

   - Experimental data file(s), containing the genes' control strengths for each sample in the experiment
     - A file listing the positive controls
     - A file listing the negative controls
     - GIF or JPEG pictures to be associated with this experiment, or with particular samples within the experiment
     - GIF or JPEG pictures of the Microarray plates the experiment was done on

   Click the **`Next`** button to proceed to the next panel. As you move to the next panel, a checkbox in the Wizard navigator will change color. You can return to any of the previous panels, by clicking the check box of the panel you would like to view again. Occasionally you will get a dialog box telling you changes in a previous panel might have detrimental effects.)

Copyright 1998-2001 Silicon Genetics

2. The Data File Format panel will appear. This panel tells GeneSpring where to look for your data files, and what kind of format they will be in. There are a number of prefabricated experiment types.

   a. Choose one of the specific types from the drop-down menu. Select **`Fully Custom`** if you are unsure which of the formats offered in the *What type of technology are you using?* box applies to you. Choosing the "Two-color experiment File" means you are using references, and the panel that asks about them will already indicate you have them. These prefabricated experiment types are included so you do not have to look at all of the possible wizard panels.

   b. At the moment, **`Locally Accessible text files`** is the only selectable option for the second drop-down menu.

   c. Click the **`Next`** button to proceed to the next panel.

3. The Properties of Experiment panel will appear.

   a. In the top box, enter the experiment name exactly as you want it to appear in the Experiments folder in the GeneSpring navigator. This name must be unique. If the name is not unique, GeneSpring will not allow you to move on to the next panel. Enter all information carefully, as GeneSpring is spelling and case sensitive.

   b. In the middle box, tell GeneSpring whether you want this experiment to appear in a subdirectory of the genome folder this experiment refers to. Clicking the **`Yes`** circle will cause another box to appear. Type in the name of any subdirectory you would like to use for this experiment. You may have more than one experiment within a folder.

   c. In the bottom box, enter any comments or general notes you have about this experiment. These notes will be visible (and editable) in the Experiment Inspector. Please refer to "Experiment and Condition Inspectors" on page 3-41 for more information about that window.

   d. Click the **`Next`** button to proceed to the next panel.

4. The Number of Arrays panel will appear. This panel tells GeneSpring how many single arrays (or samples) combine to make this experiment. A single array is defined as each time a measurement is taken of your entire set of genes.

   a. Select the **`No`** circle, if there was only a single set of measurements taken.

   OR

   a. Select the **`Yes`** circle, if more than one set of measurements for your genes were taken. Selecting Yes in this panel will reveal a box to type in the number of arrays.

   b. Enter the number of measurements that were taken of your gene set by typing the number in the *Number of Arrays* box. GeneSpring will not let you proceed if you click **`Yes`** but do not indicate how many Arrays/Samples there are.

   c. Click **`Next`** to proceed to the next panel.

 Copyright 1998-2001 Silicon Genetics

5. The Number of Parameters panel will appear. This panel tells GeneSpring how many parameters were used in this experiment, and what those parameters were. Briefly, a parameter is anything used to describe the condition or conditions of the experiment. A parameter consists of two or more parameter values; for example breast cancer, lung cancer, and healthy could be parameter values for the parameter "cancer". For a more detailed description of parameters see "Definitions of Parameters" on page 2-11.

   a. Type the number of parameters involved in this experiment in the *Number of parameters* box. Changing the number in this box changes the number of lines given in the table below.

   b. Name each of your parameters in the right-hand column (labeled *Parameter Name*). You can tab forward (or use the cursor keys in some cases) to place the cursor in the next space. When you right-click this table, there is no pop-up menu allowing you to cut and paste. You can still cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). If you right-click one of the gray areas of this table, a pop-up menu will appear.

   These pop-up menus allow you to cut and paste large sections of the table. You cannot proceed to the next panel until you have named all of your parameters. If you mis-typed the number of parameter values, just highlight over it and type in the correct number.

   c. Select the **`Next`** button to continue.

6. The Parameter Characteristics panel will appear. In this panel you can define the parameters as being numbers, plotted on a log scale, and the units associated with them.

   a. Use the scroll bars to view each parameter, selecting (by leaving a checkmark in the box) or leaving blank items for each of the parameters set up in the previous panel. You will need to type (or paste) in the units in the units box at the end of the row.
   It is perfectly acceptable to leave all the options unselected.

   b. Select the **`Next`** button to continue.

7. The How to Display the Parameters panel will appear. In this panel you tell GeneSpring what parameter types to use in the default interpretation. There are four possible choices. The default setting is **`Denotes a non-continuous variable`**, separating the data into discrete graphs viewed side by side on the screen (the non-continuous display). For more detailed information about all of these parameter displays see "Parameter Display Options" on page 2-12.

   a. Select a new option or leave the defaults for every parameter.

   b. Select the **`Next`** button to continue.

8. The Parameter Values panel will appear. In this panel you tell GeneSpring the parameter values for each condition in the experiment. Initially blank, this screen has been filled in with the Parameter Values. A parameter-value is one of the possible values a variable can have. (For a more detailed explanation of parameters and how they can be used, please see "Definitions of Parameters" on page 2-11.) In the table given in the *Parameter Values* panel, each parameter you named has its own column.

 Copyright 1998-2001 Silicon Genetics

a. You must fill in every field in each column with the appropriate parameter-value for the samples named to the far left of the field. If there are more fields than fit in the panel, scroll bars will appear. You can cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). Pasting is highly recommend because the parameter-value entries are spelling and case sensitive. If you right-click one of the gray areas of this table, a pop-up menu will appear.

The pop-up menu resulting from right-clicking the parameter labels section of the table will say copy and paste columns. The pop-up menu resulting from right-clicking the sample labels section of the table will say copy and paste rows. The pop-up menu resulting from right-clicking the gray field in the upper left-hand corner of the table will say copy and paste all. These pop-up menus allow you to cut and paste large sections of the table. Once you have filled in every field in the table you can proceed to the next panel by clicking on the **`Next`** button. If there is an unfilled box, the **`Next`** button will remain disabled.

b. Select the **`Next`** button to continue.

9. The Describe your Data Files panels will appear. This panel tells GeneSpring where to find the experimental data file pertaining to each sample. The *Describe your Data Files* panels are large. Please double-click the banner bar to expand the panel to fill your screen so you will not miss any of the possibilities.

   a. To begin describing your files to GeneSpring, you must select one of the options in the drop-down menu at the top of this panel. You have three selectable options to describe the files containing your data.

   - "All my samples are in one file"
     First and easiest, if all of your samples are in one data file select **`All my samples are in one file`**. In the table at the bottom of the panel, fill in the field labeled *File Name* with the name of the text file containing your sample's data. When your data is all in one file, the formats will all be the same. Be aware, as soon as you leave this panel, by clicking the **`Next`** button, the changes will be irrevocable. You may see the quick flutter of an error message reminding you of this.

   - "My samples are in multiple files that share a common format"
     If your samples are in different files with *exactly* the same format, select the default setting, **`My samples are in multiple files that share a common format`**.
     Enter the name of the file containing the sample data for each experiment in the table. Each file should be entered in the white boxes of the column labeled *File Name* in the same row as its sample. If your data files are where GeneSpring expects them to be (i.e., in the correct directory) the names will appear in the large white box at the bottom of the screen labeled *Files present in the current data directory.*You can double-click these names to insert those files into the *File Name* column. Each row will be filled in top-to-bottom order each time you double-click a file name until all rows are filled. If your files are not shown in the *Files present in the current data directory* box, you may not have saved your files to the correct location. If you may need to recheck the *Properties of the Experiment Set* panel. You can select from the

 Copyright 1998-2001 Silicon Genetics

list of "already viewed" panels on the left side of the Wizard to view that panel again.

If you have two files comprising a chip set you need to enter the names of both files separated by a semi-colon in the same entry blank. Please see "You might need to put more than one file in a field. To do this:" on page D-8 for more details.

Data files have the same layout when the files for each and *every* sample have exactly the same number of columns, in the same order, containing the same type of data (for example, signal intensity or background readings for the experiment). Any variation, no matter how small, means your files do not have the same layout.

If all of your sample data is in the same file, and each have the same file layout, you may need to cut and paste the information into separate files or add columns to the file you already have. For example, a data file containing the signal intensities from sample 1 and sample 2 *must* have these results in two different columns. When this is done, the control strength column in the data file pertaining to sample 1 is not in the same place as the column containing the control strength for sample 2. This means the experimental data file layout for sample 1 is *not* the same as the layout in sample 2. An experiment reported in this way, with some, but not all of the samples in the experiment reported in the same data files cannot be considered to have the same data file layout. To tell GeneSpring your data is reported in this manner, answer **`No`** to the first two questions in the *Describe your Data* panel (the *Are all of your samples in the same data file?* question, and the *Do all the data files have the same layout?* question). Enter the name of the experimental data files containing each sample in the *File Name* column of the table. Now the table allows you to repeat a file name in multiple rows (unlike the non-repetition if you answer **`Yes`** to the *Do all the data files have the same layout* question). However, if you must use the same data files the same number of times, for example sample 1-4 could be named `a.txt`, sample 5-8 could be `b.txt` and 10-12 could be `c.txt`. To continue the same example, sample 1-4 could be `a.txt`, sample 5-6 could *not* be `b.txt`, sample 7-8 could *not* be `c.txt`, and sample 10-13 could be `d.txt` as the differing numbers of samples in each file implies a different number of columns and therefor a different layout. If you have more than one data file with differing column layouts, you will have to repeat all of the subsequent panels dealing with locating which column contains what information for each data file you name.

When you right-click the table in this panel of the Experiment Wizard, there is no pop-up menu allowing you to cut and paste. You can still cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). If you right-click one of the gray areas of this table, a copy and paste pop-up menu will appear. These pop-up menus allow you to cut and paste large sections of the table. Once you have filled in every field in the table you can proceed to the next panel by clicking on the **`Next`** button.

You may see a quick flutter of an error message if GeneSpring cannot find the correct folder in your directory. Look in the TaskBar if GeneSpring will not let you go to the next panel. If an error message such as "Oops... Can't find the file:" appears use your file management system to create the correct folder and place a copy of your data file within it.

In this configuration of the *Describe your Data Files* panel, you need to click in the

 Copyright 1998-2001 Silicon Genetics

beige box in the *File Name* column, then double-click the correct file name in the *Files present* box. If the files names are not present in the box, please double check to make sure your files are saved in the correct folder within GeneSpring.

- "My samples are in multiple files with different format"
  If your samples are in various files that do not have *exactly* the same format, select **`My samples are in multiple files with different format`**. You will not be able to continue until every field is filled and GeneSpring has verified the existence of each and every file.

You might need to put more than one file in a field. To do this:

- Place one file in the field in the normal fashion.
- Manually type in a semi-colon (;) after the file name.
- Hold down the control key (Ctrl) while selecting the file you would like added to that same field.
  You can do this with either the **`My samples are in multiple files that share a common format`** option or the **`My samples are in multiple files with different format`** option.

b. Select the **`Next`** button to continue.

10. The Data File Header Lines panel will appear. The first drop-down menu in this panel allows you to tell GeneSpring whether there are any column titles in your experimental data files. If you do:

    a. Select **`has a`** line of column titles after. If you have any comment lines to discard, type the number of comment lines to be skipped the box. GeneSpring automatically skips blank lines, so you should not count blank lines among the lines to be skipped.

    b. Select the **`Next`** button to continue.

11. The Region Normalization panel will appear. This panel allows you to employ region normalizations.

    a. Select **`Yes`** at the question, *Did each of your sample(s) use multiple arrays or sections of a single array that require separate normalization?* if a sample in your experiment was preformed on more than one array, or if there is some reason you want the sections on the arrays normalized individually.

    You will need to enter the column of your experimental data file containing the region designation. Make sure the spelling and capitalization you enter is exactly the same as is used in the data file. (Copy and paste if you can to make sure the spelling and capitalization is identical.) If the region is the only entry in the region designation column, or if it is a suffix attached to the column's entry, then you need to type all of the different region designators (the different suffixes or column entries defining which gene was in which region) in the *List all possible region column entries or suffixes* box. The different region designators must be separated by spaces, or else GeneSpring will read them all as one entry.

 Copyright 1998-2001 Silicon Genetics

If the region designators used in your experimental data file are neither unique column entries, nor suffixes, see "Entering region specifications when they are not specified in their own column or as suffixes within another column" on page K-5 for how to import this information into GeneSpring. You will not be able to enter this experiment using the Wizard.

For a mathematical illustration of this normalizing option, please refer to "Normalizing Options" on page G-1.

b. Select the **Next** button to continue.

12. The Gene Name panel will appear. This panel tells GeneSpring which column of your experimental data file contains the gene names, and whether the gene name is the only entry in its column.

    a. Enter the name or number of the column containing the gene name in the box labeled *Enter the gene column name*. If you are entering the column number, count the columns from left to right, starting from one. Make sure the spelling and capitalization is perfectly consistent with your file when you are entering the column names.

    b. Select **Yes** at the second question, *Does this column contain **only** the desired gene name without suffixes or prefixes?* only if the gene name reported in the experimental data is exactly like the gene name listed in the table of genes file defining the genome.

    c. Select **Yes** in the second question if there are prefixes, suffixes, or region designators (which are frequently noted as prefixes or suffixes in the gene column).

    If you do this the next two panels presented to you will be the *Gene Name Prefix Removal* panel and the *Gene Name Suffix Removal* panel. If fewer than 10% of the gene names match your current genome, you will get a warning box.

    d. Select the **Next** button to continue.

13. The Gene Name Prefix Removal panel will appear. This panel allows you to remove one of two types of prefixes from the gene names in the experimental data file, so the gene names match the gene names given in the list of genes defining the genome. If your genes do not have prefixes it is acceptable to leave the answers to both questions **No**.

    a. If every gene has the same string of characters prepended to it, select the **Yes** circle for the first question, *Does the name appearing in the gene name column have a fixed unchanging prefix you want removed?*.

    b. Enter the string of characters prepended to your gene names in the *Enter fixed prefix* box that appears.

    Or,

    a. If every prefix is not the same for every gene it prepends, but it always ends with the same character. If this is the case, select the **Yes** circle of the second question, *Does the name appearing in the gene name column have a prefix ending in a particular character or characters?*.

    b. Enter the character marking the end of the prefix in the box labeled *Enter prefix marker character(s)**. There may be multiple different markers indicating the end of the prefix. If this

 Copyright 1998-2001 Silicon Genetics

is the case, enter them all in the **Enter prefix marker character(s)** box. Do not separate multiple markers in any way, anything you use to separate the characters, including a space, will be considered a prefix marker character and be removed from the gene name, along with anything preceding it. Make sure when you are entering a set prefix or a prefix marker character you get the spelling and capitalization exactly correct.

c. Click the **Next** button to proceed to the next panel.

14. The Gene Name Suffix Removal panel will appear. This panel allows you to remove suffixes from the gene names in the experimental data file, to make the gene names given there match the gene names given in the list of genes defining the genome. If your gene names do not have suffixes, it is acceptable to leave the answers to both questions **No**.

    If your gene names have suffixes to remove, the suffixes can be one of two types:

    a. The first is a "set" suffix; this means every gene with a suffix has the same string of characters appended to it. Click the **Yes** circle under the question *Does the name appearing in the gene name column have a fixed, unchanging suffix you want removed?*.

    b. In the box that appears, labeled *Enter suffix marker character(s)*, enter the characters of the suffix.

    Or,

    a. The other type of suffix is not the same for every gene name it appends to, but it always starts with the same character. If this is the case, select the **Yes** circle of the second question, *Does the name appearing in the gene name column have a suffix that begins in a particular character or characters?*

    b. In the box that appears, labeled *Enter suffix marker character(s)*, enter the character marking the beginning of the suffix. There may be multiple different markers indicating the beginning of a prefix. If this is the case, enter them all in the **Enter suffix marker character(s)** box. Do not separate multiple marker characters in any way. Anything you use to separate the characters, including empty spaces, will be considered a suffix marking character and will be removed from the gene name, along with any characters following it. Make sure when you are entering a set suffix or a suffix marker character you get the spelling and capitalization exact.

    c. Select the **Next** button to continue.

15. The Data Column Location panel will appear. This panel tells GeneSpring which column(s) of your experimental data files contains the genes' raw data. Enter the name or number of the column containing raw data in the *Enter data column name* box. Make sure to use the correct spelling and capitalization for this entry. If your data file includes a column containing the background signal to be subtracted from the gene's raw data, in the second question (*Do your data files contain a column representing background control strength?*) select the **Yes** circle. Enter the name of this column or its number in the white **Data Background Column** on the right. Again, beware of spelling and capitalization errors. This panel will not let you proceed to the next panel until you have entered a column name or number for the raw data column for every sample (row), and for the background column (if present).

 Copyright 1998-2001 Silicon Genetics

d. Select the **Next** button to continue.

16. The Control Channel Values panel will appear. If you have control channel values for each gene on your array then you can use this information to normalize your genes. See "Normalizing Options" on page G-1 for more information regarding how this normalization works.

    If you do not have a control for each gene (if you did a single-color experiment, this is probably the case) you should leave the **No** circle selected and proceed to the next panel.

    If you do have control channel values, select the **Yes** circle and enter the name(s) of the column (or its number) containing the control channel signals in the **Control Channel Column** box. If your experiment took a reading of the background for the control channel values, change the selection in the bottom question to **Yes**. Then, enter the column name(s) (or number(s)) of the column containing the control channel background signal. When you enter column names make sure you use the correct spelling and capitalization.

    a. Select the **Next** button to continue.

17. The Flags panel will appear. If your experimental data contains a column indicating whether the experiment worked for each gene, GeneSpring can incorporate this data. Select the **Yes** circle.

    - In the first column, enter the column name(s) (or number(s)) of the column(s) containing the pass-fail information in the **Flag column name** box.
    - In the second column, **Passed Designator**, enter the value given in the Flag column indicating the experiment worked for any particular gene. Frequently, the designator for good data is "P" for Present/Passed or "O" for OK.
    - In the third column, **Marginal Designator**, enter the value given in the Flag column indicating the experiment might have worked for any particular gene. Uncertain or marginal data is normally indicated by an "M".
    - In the fourth column, **Absent Designator**, enter the value given in the Flag column indicating the experiment did not work for any particular gene. Failed or absent data is normally indicated by an "A".

    When you are entering a column name, be sure to use the spelling and capitalization used in your experimental data file.

    If you have many rows and your designators are the same in every file click the *Guess the rest* button to fill down the table.

    a. Select the **Next** button to continue.

18. The Sample Photos panel will appear. This panel tells GeneSpring if you have any pictures you wish to associate with any or all of the samples. Pictures are nice, but they are not necessary. If you do not have any, leave the **No** circle selected and proceed to the next panel.

    If you have one or more pictures to associate with your sample, select the **Yes** circle. The panel will expand. If you have a picture already in the correct directory to associate with every sample, GeneSpring will display the file name(s) in the lower right-hand corner of the main window. In the table labeled *GIF File Name* enter the complete file name of the picture associated with the sample by double-clicking one of the file names or typing in each file name

 Copyright 1998-2001 Silicon Genetics

manually. The picture must be a *.gif* or a *.jpeg* file. If one of your samples does not have a picture associated with it, leave its field blank. GeneSpring will use the picture associated with the next closest sample.

The easiest way to fill in this table is to have all of your *.gif* or *.jpeg* files in the experiment directory. Then the file names will appear in the white box at the bottom of the panel. Just double-click on each picture in the correct order. When you right-click the *GIF File Name* table in this panel of the Experiment Wizard, there are pop-up menus allowing you to cut and paste. If you right-click one of the gray areas of this table, a pop-up menu will appear, from which you can select copy and paste options.

You can still cut and paste entries into the matrix fields by using the keyboard commands (for Windows this is Ctrl+C and Ctrl+V).

a. Select the **Next** button to continue.

19. The Array Photos panel will appear. In this panel you tell GeneSpring if you have any pictures of the array plates used. Microarray pictures are nice, but not necessary. If you don't have any, leave the **No** circle selected and proceed to the next panel. To associate Array Pictures with the samples, select the **Yes** circle for the question, *Do you have any pictures of the microarray plate(s)?* A table appears. In the *GIF File Name* column enter the complete name of the file containing the array picture to be specifically associated with the sample listed in the left-hand column. If you have an array picture for every sample GeneSpring will display it when you double-click the picture in the lower right-hand corner of the main GeneSpring window. Array pictures must be in either GIF or JPEG format.

    When you right-click the table in this panel of the Experiment Wizard, there are pop-up menus allowing you to cut and paste. You can also cut and paste entries into the matrix fields by using the keyboard commands (for Windows this is Ctrl+C and Ctrl+V).

    The pop-up menu resulting from right-clicking the **GIF File Name** label, allows you to copy and paste columns. The pop-up menu resulting from right-clicking the experiment labels section of the table, allows you to copy and paste rows. The pop-up menu resulting from right-clicking the gray field in the upper left-hand corner of the table, allows you to copy and paste all.

    These pop-up menus allow you to cut and paste large sections of the table.

    a. Select the **Next** button to continue.

20. The RT – PCR Experiments panel will appear. This panel tells GeneSpring whether the data you are loading comes from a RT-PCR experiment. RT-PCR is a technology for measuring expression levels, it reports these measurements in a different form than the standard array technologies. Instead of reporting expression values it reports:

    $$-[\log_2(\text{expression value})]$$

    If you have not dealt with RT-PCR experiments or have not heard of them before, leave the **No** circle selected, and proceed to the next panel. If you are using RT-PCR technology, select the **Yes** circle.

    a. Select the **Next** button to continue.

 Copyright 1998-2001 Silicon Genetics

21. The Normalizations: Negative Controls panel will appear. This panel tells GeneSpring if you have any genes designated as negative controls on your array, and if you want to normalize your sample using this data. You typically have negative controls when there is DNA from a different genome than the one you are investigating on the array. To indicate you have negative controls to use for normalizing, select the **Yes** circle. This normalization method takes the average signal intensities for all of the negative controls and subtracts this number from the signal intensity of each gene. For more info about this normalization option, see "Normalizing Options" on page G-1. If you do not have negative controls, or do not want to normalize your sample using the data from them select the **No** circle.

    Answering, **Yes** to the first question, *Do you have any genes designated as negative controls?* initiates a second question. If you are using negative controls you must have a file listing them, one gene name per line. This file should be in the same sub-directory as your experimental data. In the **Negative controls file name** box enter the name of the file listing your negative controls.

    For a mathematical illustration of this normalizing option, please refer to "Normalize to Negative Controls" on page G-2.

    a. Select the **Next** button to continue.

22. The Normalizations: Control Channel Values panel will appear. You will only see this panel if you have already told GeneSpring your sample has control channel values for each gene. If you have control channel values for each gene to indicate the trust you have in the experimental data for each gene, you probably want to normalize the genes by dividing their control strength by the control channel's control strength. If you have a background signal for either or both of these values, it is subtracted from the signal intensities before they are divided. For more information on this normalization option, see "Normalizing Options" on page G-1. If you wish to use this normalization, select the **Yes** circle. If you do not wish your data to be normalized using the control channel values leave the **No** circle selected.

    If you are using your control channel values for normalization, you need to enter the minimum reference signal to be used in the normalization. This is because sometimes the control channel value is very low and would artificially inflate the noise for its gene. Indicate the minimum value you would be willing to divide a gene's signal by in the *Minimum control channel strength* box.

    If you are not using your control channel values for normalization, then you are using them to indicate the trustworthiness of the experimental data for each gene. Indicate the minimum value a reference must have for you to consider the data for the gene it is associated with valid in the box labeled **Minimum confidence level**.

    For a mathematical illustration of this normalizing option, please refer to "Normalize to Control Channel Values for Each Gene" on page G-3.

    a. Select the **Next** button to continue.

23. The Normalizations: Positive Controls panel will appear. This panel tells GeneSpring if you have any genes designated as positive controls on your array and if you want to normalize your sample using this information. You typically have positive controls when there is DNA

 Copyright 1998-2001 Silicon Genetics

from a different genome than the one you are investigating on your array, and you add a known quantity of that DNA to your sample.

If you do not want to normalize your sample using positive controls leave the **No** circle selected.

a. To indicate you have positive controls for normalization, select the **Yes** circle. This normalization method takes the average signal intensities of all of the positive controls and divides each gene's signal intensity by that number. For more information about this normalization option see "Normalizing Options" on page G-1.

If you are using positive controls you must have a file specifying what the positive controls are called, listing the gene names one per line. This file should be in the same sub-directory as your experimental data. In the **Positive controls file name** box, enter the complete name of the file listing your positive controls.

Sometimes, something will go wrong with the positive controls and you will get very low values for all of them, which you will not want to use for normalization purposes. In the **Enter lower cut-off for positive controls** box, indicate the minimum average the positive controls must have such that dividing each genes' control strength by the average of the positive controls will not artificially inflate the noise of the genes. The default setting for the cut-off value is 10.

For a mathematical illustration of this normalizing option, please refer to "Normalizing Options" on page G-1.

b. Select the **Next** button to continue.

24. The Normalizations: Each Sample to Itself panel will appear. In this panel you tell GeneSpring if you want to normalize your data by making the median of all of your measurements 1, for each single sample in your experiment. (If you have not already preformed normalizations on your data you generally want to use this normalization option.) To indicate you want to normalize each sample to itself, select the **Yes** circle. Another question will appear.

    Sometimes something will go wrong with the experiment and you will get very low values for everything. In the **Enter lower cut-off value** box indicate the cut-off value. This number will be used by GeneSpring to *not* raise all of the control strength values up to a median of 1 if their average is below this number.

    For a mathematical illustration of this normalizing option, please refer to "Normalize Each Sample to Itself" on page G-6.

    a. Select the **Next** button to continue.

25. The Normalizations: Each Sample to a Hard Number panel will appear. In this panel you tell GeneSpring if you want to normalize your samples to a value you enter. You would normally only use this function if you have pre-normalized data, such as data prepared with Affymetrix's Global Scaling. In that instance, you would want to divide all data by 2500 (or whatever number you chose to normalize by in the Affymetrix software.) You will need to do this because the GeneSpring analysis algorithms assume your data is normalized to a median of 1.

    a. Select the **Next** button to continue.

 Copyright 1998-2001 Silicon Genetics

26. The Normalizations: Each Gene to Itself panel will appear. In this panel you tell GeneSpring if you want to normalize each gene to itself, so the median of all of the measurements taken for the gene is 1. If you are not doing a two-color experiment you generally want to do this, so the default setting for this panel is to perform this normalization.

    If you do not wish to employ this normalization select No radio button in the first question.

    If you wish to use this normalization, there is a second question. Sometimes something will go wrong with the experiments and all of the values for a particular gene are very low, in which case it will artificially inflate the noise of the gene if you normalize those values up to a median of 1. Specify this cut-off by entering a number in the **`Enter lower median cut-off value`** box. The default setting for the cut-off value is 0.01.

    Normalizing each gene to itself is optimal for more than five samples, as with less than five the display becomes unintuitive. Generally the better option for five samples or less is to do normalization against a particular sample.

    For a mathematical illustration of this normalizing option, please refer to "Normalizing Each Gene to Itself" on page G-8.

27. The Normalizations: All Samples to Specific Samples panel will appear. This panel tells GeneSpring if you want to normalize each sample in the experiment to a single sample within the set. Normalizing each gene to itself is often preferable to this normalization. If you wish to normalize your data in this way, select the **`Yes`** circle. Another question appears. Sometimes a gene's control strength in the sample being normalized to is anomalously low. Enter the lowest value you are willing to use for normalizations in the **`Enter lower reference cut-off value`** box.

    In the enter sample number box you can normalize multiple samples to several samples. You can also normalize several samples to several samples. You can normalize multiple samples to multiple different samples through a code like [1;2;3]1;2[3;4;5]3;4 which means normalize samples 1,2 and 3 to 1 and 2, and 4, 5 and 6 to 3 and 4. Please see "Required Syntax for Normalization to Specific Samples" on page G-10 for more information regarding the syntax to use in this panel.

    For a mathematical illustration of this normalizing option and several examples, please refer to "Normalizing All Samples to Specific Samples" on page G-10.

28. The Graphics Specifications panel will appear.

- **Defining Trust:** The upper section of this panel tells GeneSpring what the colorbar intensity scale should be, and the relative intensity values to be graphed on the y-axis in the graph display. The intensity of the colorbar in GeneSpring indicates how reliable the data for each gene is. Indicate a raw, very reliable (a high control strength) control strength value, an average (a medium control strength) value, and an unreliable (a low control strength) value in the three boxes. Any gene with a control strength above the value indicated as a high control strength will be colored using the brightest color appropriate, any gene with a control strength below the value given for unreliable data will be dull in color. The medium signal value gives the value for the mid-point of the colorbar, and genes with an average control strength are colored halfway between the two color extremes.

 Copyright 1998-2001 Silicon Genetics

For more information on how trust is expressed in the genome browser, please see the "Changing the Experimental Data Range" on page 3-36.

- **Defining default x and y values:** The middle section of the Wizard panel allows you to inspect the genes' expression profiles more closely from the genome browser. As GeneSpring does not graph the entire y-axis (the expression level axis), but only the portion most genes profiles fall into you will need to set the defaults for that portion. In the lower two boxes indicate the range of expression levels GeneSpring should graph. The values indicated here can be altered within GeneSpring (look in **`View > Change experiment interpretation`**). Here you are simply setting the defaults.
- **Defining Negative Values to Zero:** The bottom section in the Wizard panel asks if you would like to force negative values to zero. Forcing all of the negative numbers to zero converts all the negative values to zero *after* all the normalizations have been implemented and after the genes that do not pass the Pass-Fail vote have been thrown out (this happens before any normalization is applied by GeneSpring).

29. The Finish panel will appear. When you click the **`Load Now`** button all of the answers you gave in the previous Experiment Wizard panels are saved in an *.html* file.

    If GeneSpring is unable to load the data, you will get an error message with a list of the unrecognized genes that caused it not to load.

 Copyright 1998-2001 Silicon Genetics

Appendix E Installing from a Database

Custom Databases and GeneSpring

You can load experiments into GeneSpring from your company's database. To do this you will need to set up a *.database* file prior to starting the New Experiment Wizard.

Databases

A database is an organized collection of information. Essentially, it is a collection of records. In database terms, a record consists of all the useful information you can gather about a particular item. Each little bit of information making up a record is called a field. An example of a non-computerized database would be your address book. Each record represents one of your contacts, and each record consists of many fields such as name, address, number, and so on.

Computer databases automatically keep records organized and enable you to search for or pull out particular records based on any field in the record. The software allowing you to create and maintain databases is called a *Database Management System*, or DBMS. In database terminology, a *file* is called a *table*. Each *record* in the file is called a *row*, and each *field* is called a *column*.

A relational database is the most common type of database in client/server systems. Simply stated, in this type of database, relationships are established between tables based on common information.

Open Database Connectivity

Open Database Connectivity (ODBC) is an Application Programming Interface (API) allowing a programmer to abstract a program from a database. When writing code to interact with a database, you usually have to add code that talks to a particular database using a proprietary language. If you want your program to talk to Access, Fox and Oracle databases, you have to code your program with three different database languages. This can be a very difficult or time consuming task.

This is where ODBC enters the picture. When programming to interact with ODBC you only need to speak the ODBC language (a combination of ODBC API function calls and the SQL language). The ODBC Manager will figure out how to contend with the type of database you are targeting. Regardless of the database type you are using, all of your calls will be to the ODBC API. All you need to do is install an ODBC driver specific to the type of database you will be using.

Copyright 1998-2001 Silicon Genetics

Structured Query Language

Structured Query Language (SQL) is a standard language for defining and accessing relational databases. All of the major database servers used in client/server applications work with SQL. It is a query language designed to extract, organize and update information in relational databases. Each database vendor has its own particular dialect. These dialects are similar to one another, but different enough that programmers must pay close attention to which RDBMS is being used. The most important dialects of SQL are ANSI/ISO SQL, IBM DB2, SQL Server, Oracle, Ingres, and ODBC.

SQL uses statements to get work done. Examples of some of these statements are:

- SELECT
- INSERT
- DELETE
- UPDATE
- DECLARE
- OPEN
- CLOSE
- CREATE
- PREPARE
- DESCRIBE

SQL Call Level Interfaces

When a Call Level Interfaces (CLI) is used, a program requests database services by calling special SQL interface routines rather than embedding SQL statements directly into the program. There are two distinct types of CLIs. First, each DBMS vendor provides its own unique API for its database. The vendor-specific API is usually the most efficient way to access the database, but each vendor's API is unique. As a result, if you decide to write programs that use a vendor API, you lock yourself into using that vendor's DBMS. However, your programs will be efficient as possible.

The second type of CLI is a standard or open API which is supported by more than one database vendor. Several open database APIs are available, one of which is ODBC. ODBC is a standard CLI for accessing SQL databases from Windows.

The Genetic Analysis Technology Consortium

The Genetic Analysis Technology Consortium (GATC) was formed in an attempt to standardize the rapidly growing field of array-based genetic analysis. The consortium was created to provide a unified technology platform to design, process, read and analyze DNA-arrays.

The goal of the GATC is to make micro-arrays broadly available and provide a technology platform that allows investigators to use components from multiple vendors.

Copyright 1998-2001 Silicon Genetics

Databases and GeneSpring

Experimental data is not always stored on the researcher's desktop in simple text files. Sometimes the data is stored on a relational database. GeneSpring can save and load all types of data to an SQL database through ODBC.

Experimental data can be loaded from a database simply by telling GeneSpring which table(s) contain the data and which columns contain the experimental index. You then load in the data using the Experiment Wizard almost exactly as you would if they were text files (see "Entering your Prepared Database into GeneSpring" on page E-5). The only difference is you enter experiment identifiers instead of file names, and SQL table columns instead of tab-delineated column headers.

Parameters describe what the database knows about each sample. Different databases have different ways of storing parameters, so they must be retrieved by explicit SQL statements. Silicon Genetics can provide these for GATC and help write these for individual databases. This only needs to be done once. Afterwards, the customer simply chooses the database and GeneSpring will get data from it. Normalization and other options can also be set for a database.

Adding an Experiment from a Database

Make sure you have a database. Any database software can be used to produce a database. First you must make sure that GeneSpring will be able to see your database. Your database's creator should have done this already. If they have, you can skip down to "Connect your Database to GeneSpring" on page E-4.

1. Go to the control panel of your computer.
2. Select ODBC Data Sources. A new window, The OCBC Data Source Administrator, will come up.

To make a new ODBC source

1. Go to the system DSN
2. Click **Add**, which will bring up a new Create New Data Source window.
3. Select the correct type of database from the scrollable list. This will bring up a new panel.
4. Give the experiment a name. This is the name GeneSpring will use, so please remember that GeneSpring is case sensitive.
5. Click the **Select** button to browse for the correct database.
   Normally you will need to browser into a new computer (server) to access the database.
6. Now there will be a new entry in the list of databases.

Copyright 1998-2001 Silicon Genetics

Test to Make Sure Your ODBC Connection is Working

1. From Excel go to the **Data** menu.
2. Select **Get External Data**.
3. Select **New Database Query**. Look for your database in the presented list.

Connect your Database to GeneSpring

A database specification file must be set up. This is a plain text file, in a subdirectory of the main GeneSpring data directory entitled **Databases**. The text file should have the extension *.database*. This file will tell GeneSpring how to contact your database. The file contains several lines. Each line contains the name of a parameter you should set, followed by a colon, then followed by the value you want to set the parameter to.

The purpose of this file is to tell GeneSpring how to read the database as if it were a simple text file. It pulls the data together and places it in columns recognized by GeneSpring. Column names and sample name references are entered in the Experiment Wizard as normal.

1. Using your file management software, create a new folder in the data directory of GeneSpring titled **Databases**.
2. Create a file with an extension of *.database*. This file has specific requirements of what must be in it, but the items can be in any order.
   - jdbc : odbc : NameofDatabase
   - ExperimentTableName : SampleName
     If the index and gene name are separate, you will need more than one table.
     This should be a one word name. Case sensitivity depends on the database.
   - ExperimentTableIndex : which column contains the experiment number
   - GeneColumn : the column number containing the gene names
   - IntensityColumn : should contain actual results
   - debug : true
     When true it will show what commands are sent to the database when you use the Experiment Wizard.
3. Arranging your Parameters

You need to make an SQL command that will get the parameters in all samples. You can use MicroSoftQuery in Excel to generate SQL commands.

- From Excel go to the **Tools** menu.
- Select **Get External Data**.
- Select **New Database Query**.
- Make sure you tell it you want to edit in MicroSoftQuery.

Copyright 1998-2001 Silicon Genetics

GeneSpring wants:

1. Experiment ID.
2. Another experiment ID (must be unique).
3. Other parameters, Heading from tables, name of column. Double-click headings to change the name if you want.

Button at the top of the query box says SQL. Click it to get SQL statements.

SQL Get experiment and indexes : SQL statements
(this needs to be on one unbroken line, *do not* use word wrap in your text editor.

Still missing from your experiment is:

- the default normalizations
- specifications for Display Options
- specifications for Table Headings

Entering your Prepared Database into GeneSpring

Using the Experiment Wizard, select the **`Get Everything from the Database`** option.

The majority of the remaining Experiment Wizard panels will be filled in automatically.

If you left the debug setting for true an extra window will open up. When the query boxes come put these will contain actual SQL commands.

GeneSpring will have to go back to the database to get information every time you restart the program. If this takes too long, you might consider right-clicking over the correct database icon and selecting the save to disk option.

All commands in the *.experiment* files can also be added to the *.database* file.

Copyright 1998-2001 Silicon Genetics

Entering more Complicated Data from a Database

You can link various tables together in SQL. This typically requires a proficient user of databases, please check with the person who built your database if you have questions.

There are many ways to enter and organize data within databases. If the data organization in your debase if confusing, you might want to make separated tables for your data or part of your data. For example you could make a separate table just for parameters, like Table B-2.

Sample 1	Parameter Name	Parameter Value
1	elephants	2
2	elephants	34
2	daises	30

Table B-2 Sample table of mixed-up parameters

In Table B-2 you do not have parameters in the individual columns. All parameters tables should have an associated sample number somewhere.

If you use a GATC database, you will have to re-link all the sample numbers to the parameter numbers. In that case you need to define an SQL. In that case, you must define a SQL line to get those parameters, for example:

```
SQLgetParameters : select
```

This should retrieve values of and names of the parameter.

Copyright 1998-2001 Silicon Genetics

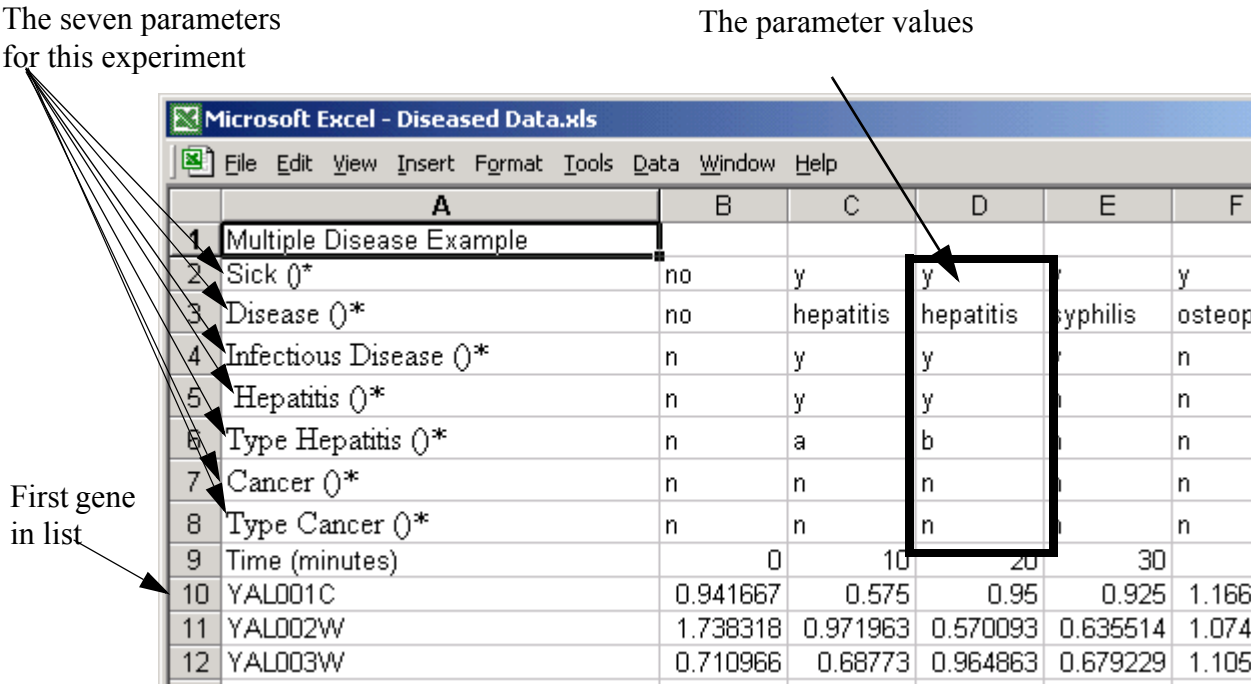
Appendix F Copying and Pasting Experiments

You can use the copy (Ctrl+C) and paste (Ctrl+V) functions to insert a new experiment or lists from the clipboard into GeneSpring. This is a very quick, but somewhat inflexible function of GeneSpring.

Preparation for Pasting

You should have normalized data in an Excel® file or saved as tab-delineated text. (Figure E-2). You must have all of the following three parts to your data. Your data *must* be in the following format to correctly paste into GeneSpring.

1. Name
   - First line must be the unique name of the experiment.
2. Parameters
   - The second line must be the first parameter (you may have as many parameters as you want, but you must have at *least* one).


The seven parameters for this experiment
The parameter values
First gene in list
Microsoft Excel - Diseased Data.xls
File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F
1	Multiple Disease Example					
2	Sick ()*	no	y	y		y
3	Disease ()*	no	hepatitis	hepatitis	syphilis	osteop
4	Infectious Disease ()*	n	y	y		n
5	Hepatitis ()*	n	y	y		n
6	Type Hepatitis ()*	n	a	b		n
7	Cancer ()*	n	n	n		n
8	Type Cancer ()*	n	n	n		n
9	Time (minutes)	0	10	20	30	
10	YAL001C	0.941667	0.575	0.95	0.925	1.166
11	YAL002W	1.738318	0.971963	0.570093	0.635514	1.074
12	YAL003W	0.710966	0.68773	0.964863	0.679229	1.105

Figure E-1 Example of parameter arrangements and values

- In the first column is the name of the parameter.
- Subsequent columns have values for parameter in that sample.
- Each parameter must have units in parentheses in the same column as the name. For example, the parameter “time” would be immediately followed by (minutes). If your parameters have no units you must follow the name with an empty set of parentheses, or GeneSpring will not recognize it as a parameter.

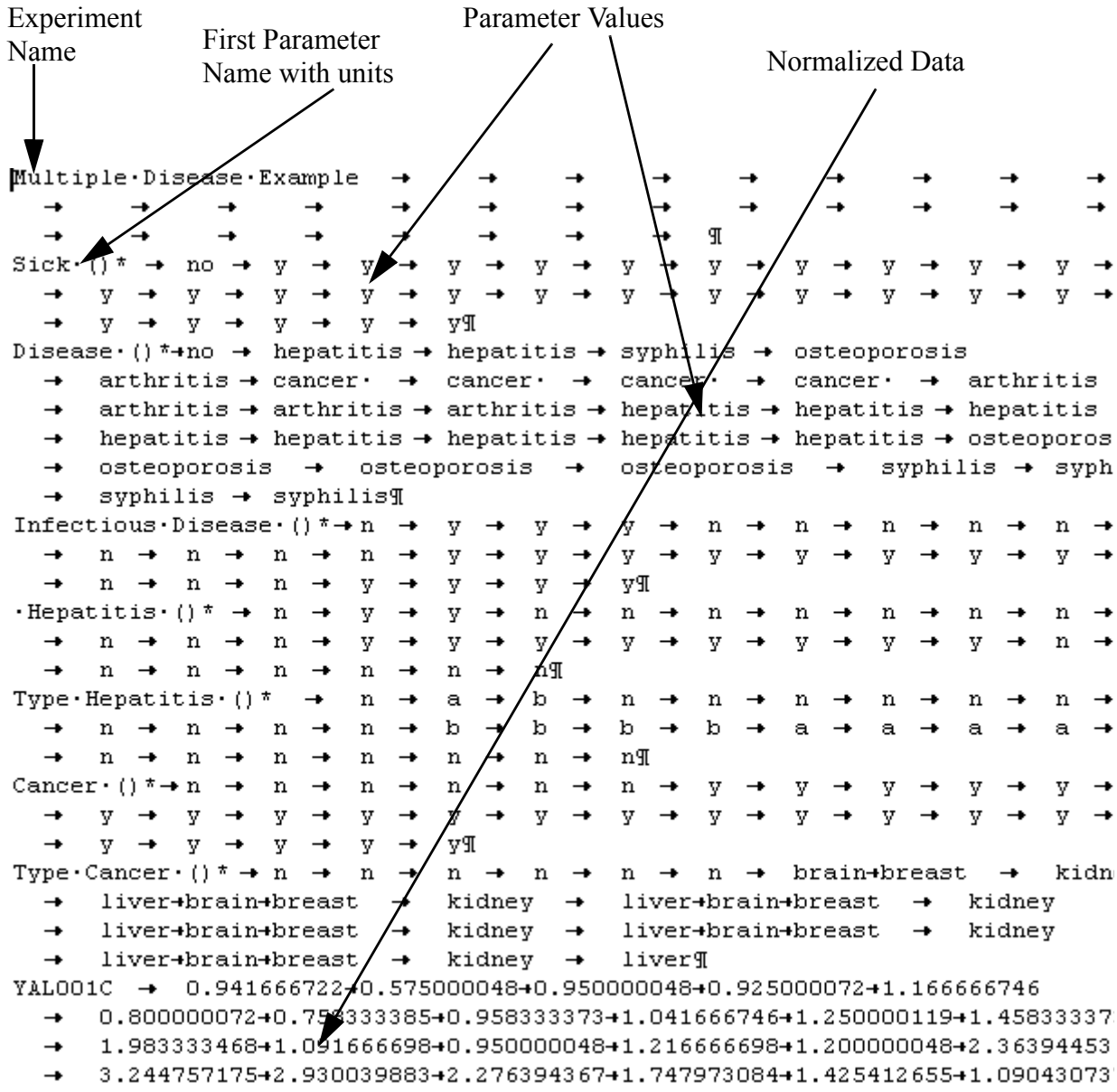
Copyright 1998-2001 Silicon Genetics

- As a default, GeneSpring assumes that the parametric values to follow are numeric and to be displayed in numerical order. If the parametric values for a parameter are non-numeric, immediately after the unit-indicating parentheses (empty if no units), enter an asterisk (*). There should be a space between right parenthesis and the asterisk (*). This tells GeneSpring to expect non-numeric parametric values and then treat the data appropriately.
- The default setting for interpretation of parameters is as a continuous element, please see "Continuous Element" on page 2-13 for details. To have the parameters treated differently, enter the following codes just after the parentheses:
- S — means the data will be interpreted as a non-continuous element, also known as a discrete element. Please see "Non-Continuous Element (Set)" on page 2-13 for details.
- C — data will be colored by the different parametric values assigned automatically by GeneSpring. In Figure E-2 each column would get a different color as time values 0-160. Please see "Color Code" on page 2-13 for details.
- R — data will be interpreted as a replicate (not shown). Please see "Replicate or Hidden Element" on page 2-13 for details.
- Of course, you can just enter all parameters with the default (no code after the parentheses) and change the interpretation later from within GeneSpring, please see "Changing the Experiment Interpretation" on page 2-17.
- For example, for the parameter `tissue type`, a non-continuous non-numeric parameter, the first column might look like:
  `tissue type() *S`.

If you have no parameters give it arbitrary (but meaningful) names so you will be able to distinguish each sample from those in other columns.

3. Data
   - There can only be one gene per line.
   - The name of gene must be in the first column.
   - The following columns are data points for each parameter.

Copyright 1998-2001 Silicon Genetics




```
Multiple·Disease·Example  →     →     →     →     →     →     →     →     →
  →     →     →     →     →     →     →     →     →     →     →     →     →
  →     →     →     →     →     →     →     →  ¶
Sick·()* → no → y → y → y → y → y → y → y → y → y → y → y → y →
  → y → y → y → y → y → y → y → y → y → y → y → y →
  → y → y → y → y → y¶
Disease·()*→no → hepatitis → hepatitis → syphilis → osteoporosis
  → arthritis → cancer· → cancer· → cancer· → cancer· → arthritis
  → arthritis → arthritis → arthritis → hepatitis → hepatitis → hepatitis
  → hepatitis → hepatitis → hepatitis → hepatitis → hepatitis → osteoporos
  → osteoporosis → osteoporosis → osteoporosis → syphilis → syph
  → syphilis → syphilis¶
Infectious·Disease·()*→n → y → y → y → n → n → n → n → n →
  → n → n → n → n → y → y → y → y → y → y → y → y →
  → n → n → n → y → y → y → y¶
·Hepatitis·()* → n → y → y → n → n → n → n → n → n → n →
  → n → n → n → y → y → y → y → y → y → y → y → n →
  → n → n → n → n → n → n¶
Type·Hepatitis·()* → n → a → b → n → n → n → n → n → n →
  → n → n → n → n → b → b → b → b → a → a → a → a →
  → n → n → n → n → n → n → n¶
Cancer·()*→n → n → n → n → n → n → y → y → y → y → y →
  → y → y → y → y → y → y → y → y → y → y → y → y →
  → y → y → y → y → y¶
Type·Cancer·()* → n → n → n → n → n → n → brain→breast → kidn
  → liver→brain→breast → kidney → liver→brain→breast → kidney
  → liver→brain→breast → kidney → liver→brain→breast → kidney
  → liver→brain→breast → kidney → liver¶
YAL001C → 0.941666722→0.575000048→0.950000048→0.925000072→1.166666746
  → 0.800000072→0.758333385→0.958333373→1.041666746→1.250000119→1.45833337
  → 1.983333468→1.091666698→0.950000048→1.216666698→1.200000048→2.36394453
  → 3.244757175→2.930039883→2.276394367→1.747973084→1.425412655→1.09043073
```

Figure E-2 Example of a correctly formatted tab-delineated file

Most Common Mistakes in Pasting

- forgetting the title
- not using parentheses
- not having parameters
- using unnormalized data
- having extraneous columns
- forgetting to indicate parameters having non-numeric parametric values with an asterisk (*)

Copyright 1998-2001 Silicon Genetics

- using more than one type of decimal marker, or the wrong type for your computer's settings (If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably. For details on changing the language settings in GeneSpring, please refer to "The Miscellaneous" on page B-5.)

Pasting your Experiment into GeneSpring

If you have not already, give your experiment a unique name. If it turns out it is not a *unique* name, then GeneSpring will append a number on the end to distinguish it from other experiments of the same name.

You can copy (Ctrl+C) all or part of a correctly set up Excel® or tab-delineated file.

In the main GeneSpring window, go to **`Edit > Paste > Paste Experiment`**.

GeneSpring will automatically update the window, regardless of which display options you currently have active.

Larger files may take longer to paste, depending on your system.

WARNING! Some computers will have a limit on the amount of data you can place on the clipboard. If you are consistently crashing at the point, you may need a Java virtual machine update.

GeneSpring will bring up a new Choose Experiment Name box, with the current name of the experiment already in the Name text box.

GeneSpring will take you back to main window with your new experiment already on display. From here, you can alter the normalizations with **`Experiment > Change Normalizations`** command or alter the interpretation with the **`Experiment > Change Interpretation`** command.

Copying an Experiment or a List Out of GeneSpring

Choose an experiment or a gene list from the navigator. When you choose to copy and experiment, please be aware you will copy only the gene list currently selected. If you want to copy *all* the genes in your currently-viewed experiment, please right-click over the "All genes" list and select **`Display List`** before you begin to copy.

In the main GeneSpring window, select **`Edit > Copy > Copy Experiment`**.

Your data will be saved to the clipboard. From there you can paste your experiment or gene list into Microsoft® Notepad, Microsoft® Word or Microsoft® Excel.

When you paste the gene list will be sorted into the order presented in the Ordered List view.

 Copyright 1998-2001 Silicon Genetics

Appendix G Normalizing Options

To normalize in the context of DNA microarrrays means to standardize your data too be able to differentiate between real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales your data so that you can compare relative gene expression levels. GeneSpring offers the following normalization options :

There are several normalization options available in GeneSpring:

- Normalize to Negative Controls also referred to as Background Subtraction
- Normalize to Control Channel Values for Each Gene, also referred to as Per Spot normalization
- Normalize to Positive Controls
- Normalize Each Sample to Itself, also referred to as Normalizing to the distribution of all genes
- Normalizing to a Constant Value (hard number)
- Normalize Each Gene to Itself, also referred to as Normalizing to the median for each gene
- Normalize all Samples to Specific Sample
- Region Normalization

You can follow the directions in any or all of these sections, as appropriate, to normalize your data. In a few cases, it would not make sense to apply two options together, for instance: normalizing each sample both to a positive control and across the whole sample, or normalizing each gene to itself (across all samples) and to a specific sample. The GeneSpring Experiment Wizard will only allow you to choose one of each of these. Other than those instances, you may choose any options appropriate to your data. The order the normalizations are performed in is mathematically significant. GeneSpring performs normalizations in the order listed above. Three normalizations can be applied either to samples or regions (normalize to negative controls, normalize to positive controls, and normalize each sample or region to itself) and are assumed to apply to samples unless otherwise specified. See “Region Normalization” on page G-15 for further information. For instructions on how to implement any of these normalizations from within GeneSpring, see “Experiment Normalizations” on page 2-21. There is one normalization in addition to those listed whose implementation is automatic: repeated measurements in a single data file are assumed to be repeats and will be averaged before any of the six main normalizations are implemented. See “Dealing with Repeated Measurements” on page -16 for details.

Copyright 1998-2001 Silicon Genetics

Background Subtractions

When considering how to transform raw data to normalized data, the first thing that may be necessary is to subtract an estimate of background level. The background level is taken from a separate column in your data set. Typically there will be a column labeled *negative control* containing information on the background level data. The median value of the negative controls will be subtracted from the raw values for each gene before anything else is done.

Normalize to Negative Controls

If you have any genes designated as negative controls on your array (usually, you have negative controls when there is DNA from a different genome than the one you are investigating on the array), you can normalize the data using this information. This normalization removes the background from the experimental readings by giving you a general idea of the lowest amount of exposure possible for signals taken from a particular array and then subtracting this amount from your raw experimental results. The formula used is:

(the control strength of gene A in sample X)
-(the median signal of the negative controls in sample X)

Once you normalize to negative controls, you probably want to either normalize to positive controls or each sample to itself and then normalize each gene to itself.

Mathematical Illustration of the Normalize to Negative Controls Method

Given the raw data with negative controls:

Raw Experimental Results			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1008	2060	1510
CLN2	1008	2060	510
CDC28	108	260	60
HSL1	1008	2060	510
YGP1	10 008	20 060	5010
Control 1	7	58	10
Control 2	8	60	0
Control 3	9	63	20

Copyright 1998-2001 Silicon Genetics

The same data normalized to negative controls:

After Normalizing to Negative			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1000	2000	1500
CLN2	1000	2000	500
CDC28	100	200	50
HSL1	1000	2000	500
YGP1	10 000	20 000	5000
Median of the Controls	8	60	10

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

Normalize to Control Channel Values for Each Gene

Control Channel Values are intended to provide a baseline. Different samples can be compared to the baseline and to one another. By using these comparisons, you can determine variations caused by the particular experimental conditions you are exploring, rather than the overall sample conditions. If you have a control channel value to indicate the trust you have in your experimental data, you probably want to normalize the genes by dividing their signal strength by the control’s signal strength. The formula for this normalization option looks like this:

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(control channel value for gene A in sample X)}}$$

In two-color experiments the control channel is often a green signal. If you normalize to the control channel for each gene you may also want to normalize each sample to itself or to a positive control. This will provide a control for sources of variability affecting the whole chip, for example, variations in the amounts of dye added, etc. You probably do not, however, need to normalize each gene to itself or to a single control sample.

Copyright 1998-2001 Silicon Genetics

Mathematical Illustration of the Normalize to a Control Channel Value for Each Gene Method

Given raw data with a Control Channel:

Raw Experimental Results						
Gene Name	Sample 1	Reference 1	Sample 2	Reference 2	Sample 3	Reference 3
CLN 1	1000	1000	2000	2000	1500	500
CLN2	1000	1000	2000	2000	500	500
CDC28	100	100	200	200	50	50
HSL1	1000	1000	2000	2000	500	500
YGP1	10 000	10 000	20 000	20 000	5000	5000

The results of normalizing to a control channel for each gene:

After Normalizing to a Control Channel Value for Each Gene			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1	1	3
CLN2	1	1	1
CDC28	1	1	1
HSL1	1	1	1
YGP1	1	1	1

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

 Copyright 1998-2001 Silicon Genetics

Normalize to Positive Controls

This normalization method is intended to remove the differences in amount of exposure between samples, providing you with a baseline so different samples are comparable to one another. Positive controls give you a general idea of how well the array responded to exposure. Normalizing to positive controls will factor in this information with the experimental results you analyze. You can normalize your data with this method if you have genes designated as positive controls on your array (you usually have positive controls when there is DNA from a different genome than the one you are investigating on your array, and you added a known quantity of that DNA to your sample). The formula used to do this is:

$$\frac{\textbf{(the signal strength of gene A in sample X)}}{\textbf{(the median signal of the positive controls in sample X)}}$$

This normalization should not be used with normalizing each sample to itself, as they are both intended to address the same issue. After normalizing to positive controls you probably still want to normalize each gene to itself.

Mathematical Illustration the Normalize to Positive Controls Method

Given raw data with positive controls:

Raw Experimental Results			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1000	2000	1500
CLN2	1000	2000	500
CDC28	100	200	50
HSL1	1000	2000	500
YGP1	10 000	20 000	5000
Control 1	5000	10 000	2500
Control 3	2000	4000	1000

The results of normalizing to positive controls:

After Normalizing to Positive Controls			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	0.5	0.5	1.5
CLN2	0.5	0.5	0.5
CDC28	0.05	0.05	0.05

Copyright 1998-2001 Silicon Genetics

After Normalizing to Positive Controls			
Gene Name	Sample 1	Sample 2	Sample 3
HSL1	0.5	0.5	0.5
YGP1	5	5	5

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

Normalize Each Sample to Itself

This normalization is intended to remove the differences in amount of exposure between samples, so different samples are comparable to one another. This method makes the median of all of your measurements 1, for each sample. The formula used to do this is:

$$\frac{\text{(the signal strength of gene A in sample X)}}{\text{(the median of all of the measurements taken in sample X)}}$$

This normalization should not be used with normalizing to positive controls, as they are both intended to address the same issue. If you do not have either positive controls or a reference it is strongly suggested you normalize each sample to itself.

This option is also referred to as Distribution of All Genes or Global Scaling. Please refer to “Normalizing to the Distribution of All Genes” on page 2-23 and “Negative Control Strengths” on page G-18.

Mathematical Illustration of the Normalize Each Sample to Itself Method

Given raw data without positive controls or control channel:

Raw Experimental Results			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1000	2000	1500
CLN2	1000	2000	500
CDC28	100	200	50
HSL1	1000	2000	500
YGP1	10 000	20 000	5000

 Copyright 1998-2001 Silicon Genetics

The results of normalizing each sample to itself:

After Normalizing Each Sample to Itself			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1	1	3
CLN2	1	1	1
CDC28	0.1	0.1	0.1
HSL1	1	1	1
YGP1	10	10	10

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

Normalizing Each Sample to a Hard Number

You would normally only use this function if you have pre-normalized data, such as data prepared with Affymetrix’s Global Scaling™. In that instance, you would want to divide all data by 2500 (or whatever number you chose to normalize by using the Affymetrix software). You will need to do this because the GeneSpring analysis algorithms assume your data is normalized to a median of 1. GeneSpring will use the following formula:

$$\frac{\textbf{(the signal strength of gene A in sample X)}}{\textbf{(hard number in sample X)}}$$

You can use this normalization in concert with Normalize Each Gene to Itself.

Please refer to section “The Normalizations: Each Sample to a Hard Number panel will appear. In this panel you tell GeneSpring if you want to normalize your samples to a value you enter. You would normally only use this function if you have pre-normalized data, such as data prepared with Affymetrix’s Global Scaling. In that instance, you would want to divide all data by 2500 (or whatever number you chose to normalize by in the Affymetrix software.) You will need to do this because the GeneSpring analysis algorithms assume your data is normalized to a median of 1.” on page -14 or to the “Use Constant Values” on page 2-24 for more details.

Copyright 1998-2001 Silicon Genetics

Normalizing Each Gene to Itself

This normalization method is intended to remove the differing intensity scales from multiple experimental readings. It normalizes each gene to itself, so the median of all of the measurements taken for that gene is one. With this normalization, you may graph a set of similar genes (defined as similar by using the correlation coefficient) and the experimental points will be graphically similar to one another. They are all on the same vertical scale, rather than the same pattern of changes on widely differing vertical levels. The formula used is:

$$\frac{(\textbf{the signal strength of gene A in sample X})}{(\textbf{the median of every measurment taken for gene A throughout all of the samples})}$$

Do not use this normalization method in concert with normalizing all samples to a specific sample, as they are both intended to address the same issue. If you are using GeneSpring to do all of your normalizations, and you are not doing a two-color experiment, using this normalization method is highly recommended. This normalization option is commonly combined with either normalizing each sample to itself or normalizing to positive controls. As it is more striking mathematically to illustrate it as the second step of normalization, there are two mathematical illustrations, one following the normalization of each sample to itself, and the second following normalization to positive controls. For explanations of either of these first normalizations see “Normalize Each Sample to Itself” on page -6 or “Normalize to Positive Controls” on page -5.

You can specify a cutoff to prevent small and negative measurements from participating in the normalization. The cutoff is specified in terms of measurement values that have been partially normalized in previous normalization steps, so if your data has other (e.g. per-sample) normalizations, this should probably be a small number, like 0.01.

Obviously, this normalization needs more than one sample to make sense. It can be considered a synthetic control.

Mathematical Illustration of the Normalizing Each Gene to Itself Method

Data normalized by *Normalize Each Sample To Itself*:

After Normalizing Each Sample to Itself			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1	1	3
CLN2	1	1	1
CDC28	0.1	0.1	0.1
HSL1	1	1	1
YGP1	10	10	10

 Copyright 1998-2001 Silicon Genetics

The results of normalizing each gene to itself:

After Normalizing Each Gene to Itself			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1	1	3
CLN2	1	1	1
CDC28	1	1	1
HSL1	1	1	1
YGP1	1	1	1

Data normalized by *Normalize to Positive Controls*:

After Normalizing to Positive Controls			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	0.5	0.5	1.5
CLN2	0.5	0.5	0.5
CDC28	0.05	0.05	0.05
HSL1	0.5	0.5	0.5
YGP1	5	5	5

The results of normalizing each gene to itself:

After Normalizing Each Gene to Itself			
Gene Name	Sample 1	Sample 2	Sample 3
CLN 1	1	1	3
CLN2	1	1	1
CDC28	1	1	1
HSL1	1	1	1
YGP1	1	1	1

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

 Copyright 1998-2001 Silicon Genetics

Normalizing All Samples to Specific Samples

This normalization option is intended to remove differing intensity scales from each sample by comparing all of the samples to one or more specific samples. The formula for this is:

$$\frac{\text{(the signal strength of gene A in sample X)}}{\text{(the signal strength of gene A in the control sample(s))}}$$

Do not use this normalization method in concert with normalizing each gene to itself or normalizing to control channel values, as they are all intended to address the same issue. Unless your experiment was designed with specific control samples, it is recommended you normalize each gene to itself (i.e. to the median across all samples) rather than using this normalization method.

Only use this normalization if you have control samples for which you consider the measurements very reliable and you want all of the measurements for the other samples to be in relation to those very reliable samples. You will need normalization definitions for *all* your samples before you begin this.

Required Syntax for Normalization to Specific Samples

In this scenario you will need to use a very specific syntax to describe your samples.

If you are normalizing to a single sample, indicate the sample number in the box labeled **`Enter Sample Number(s)`**.

If you wish to normalize all of your samples to the mean of a set of control samples, indicate the sample numbers of the control samples. Multiple sample numbers must be separated by commas (e.g. 1,2). Ranges of sample numbers can be indicated by a dash (e.g.1-3,5).

- Example 1:

  1-3,5

  Translation: normalize all samples to the mean of samples 1, 2, 3, and 5.

Alternatively, you can normalize subsets of samples to the mean of specific subsets of control samples. Begin by listing those samples to be used as controls for a majority of the samples (as described above). For samples to be normalized to the mean of a different set of samples, add (in parentheses) a list of sample numbers for the samples to be normalized, followed by a colon, followed by a list of sample numbers for the control samples. You may specify as many of these lists as you need.

- Example 2

  1(5:4)

  Translation: normalize all samples to sample 1 (including sample 4), except for sample 5, which should be normalized to sample 4.

Copyright 1998-2001 Silicon Genetics

- Example 3

  1(5,6:4)(7-10:7,8)

  Translation: normalize all samples to sample 1 except for samples 5, 6, and 7 through 10. Sample 5 and 6 should be normalized to sample 4, and sample 7 through 10 should be normalized to the mean of samples 7 and 8.

- Example 4

  1,2(3-5,7:3-4)(6,8-9:5)

  Translation: all samples will be normalized to the arithmetic mean of samples 1 and 2, except for samples 3 through 5, and 7, which will be normalized to the average of samples 3 and 4. In addition, samples 6, 8, and 9 will be normalized to sample number 5.

- Example 5

  The various parenthetical phrases will occur all at once, so you may place any piece in any place in the string.

  (1,2:7)(7:7)(3,4:8)(8:8)(5,6:9)(9:9) is the same as

  (7:7)(1,2:7)(8:8)(3,4:8)(5,6,9:9) is the same as

  (1,2,7:7)(3,4,8:8)(5,6,9:9) is the same as

  7(3,4,8:8)(5,6,9:9)

  Translation: samples 1, 2, and 7 will be normalized to sample 7, samples 3, 4, and 8 will be normalized to sample 8, and samples 5,6, and 9 will be normalized to sample 9. All values for the normalized samples 7, 8, and 9 will equal one.

If you have a cutoff, then the scaling factor for this step of the normalization is computed by taking the arithmetic mean over the set of control sample measurements that have values (are not N/A) and are above the cutoff. If no such values are present for a given gene, then a special normalization is done. In this case, the cutoff value itself is used as the basis of the normalization. Any sample with a measurement level greater than or equal to the cutoff will be normalized by this factor, and any sample with measurement level less than this cutoff will be have a normalized value set to N/A. This is done in order to avoid losing data for genes that might have low measurement levels in the control group, but significantly upregulated levels in the treatment groups, without introducing artificially downregulated values.

Copyright 1998-2001 Silicon Genetics

Special cases

As an example, you might have patients, controls and drugs arranged in the following manner. There are a total of nine samples.

	Control	Patients	
Drug X	7	1	2
Drug Y	8	3	4
Drug Z	9	5	6

- To normalize the control to itself, use this syntax:

  (1,2,7:7)(3,4,8:8)(5,6,9:9)

  This will finish with sample 1 divided by raw 7, 2 divided by raw 7 and 7 divided by raw 7. All values for the normalized sample 7 will equal one.

- To normalize the control to the average of controls:

  If you want to see sample 1 divided by the raw 7, sample 2 divided by raw 7 and sample 7 divided by the average of 7, 8 and 9, you must use this syntax:

  (1,2:7)(3,4:8)(5,6:9)(7,8,9:7,8,9)

  This will divide sample 1 by the raw data of 7, sample 2 by the raw data of 7 and sample 7 by the average of sample 7, 8 and 9.

Mathematical Illustration of the Normalizing Samples to a Specific Sample Method

As an example, your experiment might be designed with three different types of tissues, 3 control samples and 6 treated samples arranged in the following manner. There are a total of nine samples.

	Control	Treated	
Tissue Type X	Sample 7	Sample 1	Sample2
Tissue Type Y	Sample 8	Sample 3	Sample 4
Tissue Type Z	Sample 9	Sample5	Sample 6

Copyright 1998-2001 Silicon Genetics

The results of normalizing each sample to itself:

After Normalizing Each Sample to Itself									
	Treated Samples						Controls		
	Tissue X		Tissue Y		Tissue Z		Tissue X	Tissue Y	Tissue Z
Gene Name	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9
CLN 1	1	1	2.5	3	1.5	1.5	1	1	1.5
CLN2	1	1	1	1	1	1	1	1	1
CDC28	0.1	0.1	0.5	0.5	0.5	0.1	0.1	0.5	1
HSL1	1	1	4	4	2	2	1	4	2
YGP1	15	10	20	20	10	10	10	20	10

Samples 1, 2 and 7 are normalized to sample 7, and samples 3, 4, and 8 are normalized to sample 8, and samples 5, 6, and 9 are normalized to sample 9. Note that the normalized data for every gene in each of the three control samples will be 1.

Copyright 1998-2001 Silicon Genetics

After Normalizing Each Sample to the Control Sample									
	Treated Samples						Controls		
	Tissue X		Tissue Y		Tissue Z		Tissue X	Tissue Y	Tissue Z
Gene Name	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9
CLN 1	1	1	2.5	3	1	1	1	1	1
CLN2	1	1	1	1	1	1	1	1	1
CDC28	1	1	1	1	0.5	.1	1	1	1
HSL1	1	1	1	1	1	1	1	1	1
YGP1	1.5	1	1	1	1	1	1	1	1

Another way to use this normalization method requires that your experiment be designed to have a set of controls that you wish to use, en mass, as the controls for your experiment. In other words, you want to normalize all of your samples to the arithmetic mean of a set of controls.

After Normalizing Each Sample to Itself						
	Treated Samples			Controls		
Gene Name	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6
CLN 1	1	1	3	1	1	1
CLN2	1	1	1	0.5	1	1.5
CDC28	0.1	0.1	0.1	0.1	0.1	0.1
HSL1	2	2	2	0.5	0.5	5
YGP1	10	10	10	10	10	10

Copyright 1998-2001 Silicon Genetics

After normalizing each sample to itself the samples are normalized to samples to the average of the controls. Note that this allows you to analyze the variability among the controls as well as the treated samples.

After Normalizing All Samples to the Average of the Controls						
	Treated Samples			Controls		
Gene Name	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6
CLN 1	1	1	3	1	1	1
CLN2	1	1	1	0.5	1	1.5
CDC28	1	1	1	1	1	1
HSL1	1	1	1	.25	.25	2.5
YGP1	1	1	1	1	1	1

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring.

Region Normalization

This normalization option allows you to normalize sections of a sample rather than normalizing over the entire sample. This is especially important if you used multiple arrays for each experimental point or if there is some reason you need to normalize sections of an array separately from one another. Region normalization is not a separate mathematical formula the way the previous normalizations discussed in this chapter are. Using this normalization means if you normalize to negative controls, to positive controls or normalize each sample to itself you do not actually normalize over each sample, but rather perform the normalization over each region. Hence the formulas for these three normalization options become:

Normalizing to Negative Controls for a Region:

(the control strength of gene A in region Y of sample X)
-(the median signal of the negative controls in region Y of sample X)

Normalizing to Positive Controls for a Region:

(the control strength of gene A in region Y of sample X)
(the median signal of the positive controls in region Y of sample X)

Normalizing Each Region to Itself:

(the control strength of gene A in region Y of sample X)
(the median of all of the measurements taken in region Y of sample X)

See “Experiment Normalizations” on page 2-21 for how to implement this normalization option from within GeneSpring and for how to define a region.

 Copyright 1998-2001 Silicon Genetics

Dealing with Repeated Measurements

Single Data File

Occasionally the raw experimental data in the data file for your sample will have more than one line devoted to a particular gene. This may be because you did the sample twice or because you did the sample once but took the measurements twice. If the same gene name is reported multiple times on different horizontal lines in your data file, GeneSpring will automatically consider the measurements repeats and average all of the control strengths together.
GeneSpring will report the average to you, and it will keep track of the minimum and maximum values for each gene, but GeneSpring will not be able to access the particular values falling between the minimum and maximum values. The formula for averaging a repeated gene is:

$$\frac{[(\textbf{the signal strength of gene A1}) + (\textbf{the signal strength of gene A2}) + \ldots + (\textbf{the signal strength of gene An})]}{\textbf{N}}$$

This process is done for every gene repeated in a data file, and it is done *before* any other normalizations are applied to the raw values.

Frequently samples are repeated with exactly the same parameters, but are reported in different data files. If this is the case, the fact the samples are repeats is represented via parameter. The same normalization is employed when dealing with an experimental parameter considered to be a repeat, but in that case the averaging takes place after the raw data for each gene has been normalized. See “Change Experiment Parameters” on page 2-8 for more information about repeats reported in separate data files.

Mathematical Illustration of the Dealing with Repeated Measurements in a Single Data File Method

Given this raw data, with four repeats of YMRI99W (marked with the arrows):

➢	YMR199W	1117
	YMR200W	1384
	YMR201C	1101
	YMR202W	1357
	YMR203W	1162
	YMR204C	1464
	YMR206W	978
	YMRI99W	973
	YMR207C	1618
	YMR208W	1374
	YMR209C	1432
	YMR210W	1068
	YMR211W	1568
➢	YMR199W	1313
	YMR212C	1638
	YMR213W	1648
	YMR214W	1282
➢	YMR199W	1218
➢	YMR199W	1496

GeneSpring averages all of the measurements of YMR199W to get an average control strength of 1286. GeneSpring notices the maximum control strength for YMR199W in this sample is 1496

 Copyright 1998-2001 Silicon Genetics

and the minimum is 1117. These values are the end points of YMR199W's error bar which GeneSpring will plot when you choose to display error bars in either the graph or the scatter plot displays. After this average has been taken, GeneSpring discards any measurements between the end points. Hence the measurements 1313 and 1218 will be automatically discarded.

Measurement Flags

Measurement flags are markers in your data set indicating whether or not any given measurement is regarded as "Passed (or OK)", "Marginal", "Absent" or "Failed". Data is assigned one of four flags.

Flags assigned by you when the experiment in entered into GeneSpring:

- **Good Data:** data is present and reliable. Marked with a "P" for passed or "O" for ok.
- **Marginal Data:** data is present, but of unknown or dubious quality. Marked with an "M" for marginal.
- **Absent Data:** there is no data available, and there should have been. Marked with an "A" for absent or "F" for failed.

Flags assigned by GeneSpring:

- **Unavailable Data:** if there is no flag in the column, GeneSpring will assign that measurement a "U".

Only measurements at the "highest" available level of flag are combined and treated as replicates in GeneSpring version 4.0. The order of flag precedence is P M U A. If one or more Ps are present, only Ps are used, if not, and one or more Ms are present, then only Ms are used, etc. Summary statistics are collected over these cases and stored, with the corresponding flag. All other flag data is discarded for the gene. This is done when the experiment is loaded into GeneSpring and is not affected in any way by later user choices about which codes are to be used or displayed. The only way to avoid this is to not declare a flag column during data load, which means that the flags would not be available for other uses.

For information about measurement flags and how to load them into your experiment, please refer to "The Flags panel will appear. If your experimental data contains a column indicating whether the experiment worked for each gene, GeneSpring can incorporate this data. Select the Yes circle." on page D-11 and "Measurement Flags" on page J-12.

 Copyright 1998-2001 Silicon Genetics

Negative Control Strengths

Some types of microarray technology report negative control strengths. This is usually the result of subtracting estimated background levels that are larger than the raw signal. This can happen in situations where the expression levels of the gene are low compared to the measurement error. It can also happen when there is background subtraction or when a mismatched probe set has higher intensity levels than the perfect match probe sets.

If negative signal levels occur in a large fraction of the data used for normalization, there can be problems with the normalization, as the median across the normalization set can be very small or even negative. This leads to unreasonable results of normalization. In such cases, which only occur in a few situations, GeneSpring does an extra step in the normalization, where it readjusts the background level for that data by adding a constant to all the raw control strengths in such a way that the 10th percentile of the signal is set equal to 0, before proceeding with the median normalization. This correction, called the affine background correction, is applied only when the 10th percentile of the data is more negative than the median of the data is positive. You will get a warning message when you first load your data into GeneSpring if this background correction has been applied. Also, in the Gene Inspector raw control strengths adjusted by this correction are flagged with asterisks.

Whether or not the above correction is applied, negative signal levels may still be present for a few measurements. GeneSpring offers the option as the last step of normalization to set these values to zero. Also, when interpreting data in logarithm or fold interpretations, GeneSpring treats all normalized ratio values less than 0.01 (including 0 and negative values) as if they had a ratio of 0.01 preventing transformation problems.

Normalization for Particular Array Types

For Affymetrix or One-color experiments, you should normalize each sample to itself (as described in “Normalize Each Sample to Itself” on page -6) *and* normalize to a single sample” (as described in “Normalizing All Samples to Specific Samples” on page -10). Or, you can normalize each gene to itself (as described in “Normalizing Each Gene to Itself” on page -8).

For Two-color experiments, normalize each gene to reference (as described in “Normalize to Control Channel Values for Each Gene” on page -3). Then, normalize each sample to itself (as described in “Normalize Each Sample to Itself” on page -6), that is not done by your scanner software.

 Copyright 1998-2001 Silicon Genetics

Appendix H Creating Folders for New Genomes

Normally, GeneSpring will create new folders for you when you use the Genome Wizard. See “Genome Wizard” on page C-1 for more details.

To manually create a new folder in the genome browser, you must go through a file management system, such as Windows Explorer®. For example, a new folder named “Mouse” has would be created and placed into the data directory of GeneSpring.

Before your new Mouse folder will appear in GeneSpring navigator you will need to create a correct *mouse.genomedef* file. A *.genomedef* file will contain all the information GeneSpring needs to create a folder and other data objects. Make sure you save the *.genomedef* file in the correct folder (the “Mouse” folder) after you create it. Please see “The .genomedef File” on page I-1 for details on creation.

Raw Data

What Data Are Necessary?

You must have a list of distinct names for all the genes you intend to work with. In addition, a genome may also have GenBank Accession Number, sequences, alternative names, functional information, map positions, EC numbers, and so on, associated with genes. It may also include links to web-based databases. Each genome should have a distinct name, to reduce confusion.

What Format do these Data Need to be in?

Your Master Gene Table file

You will generally need either a Master Gene Table or a GenBank/EMBL entry for your organism. If you use a Table of Genes containing the genes’ GenBank Accession Numbers, then the GenBank information associated with each gene can be automatically updated. See “Updating your Master Gene Table with GeneSpider” on page 2-15 for how to do this.

There are four possible formats for a Master Gene Table: **“name list”**, **“name function”**, **“SGD”**, and **“Mapped”**. The reason these formats are called *Master Gene Table* is because it is easiest to create them in spreadsheet programs, such as Microsoft Excel®, and then use the **Save As** command to create tab-delineated text files. Occasionally a Master Gene Table is referred to as the Table of Genes, the Master Gene List or the Array Element List.

Name List

The simplest format for a Master Gene Table is **“name list”**. In this format the Master Gene Table is a single column comprised of the names of the genes:

```
Gene1
Gene2
Gene3
```

 Copyright 1998-2001 Silicon Genetics

Gene names with spaces in them, such as “Gene 1” are acceptable.

Name Function

The next simplest format for the Master Gene Table is **“name function”**. In this format the table of genes is the same as the table for “name list” except each gene may be followed by a description of its function. If you have additional information about the genes, enter it in the same row as the gene it refers to, separated from the gene name by a tab character or column separator in Microsoft Excel®. An example of this is:

```
Gene1   Putative Phosphokinase
Gene2
Gene3   Deletion causes 2 tails
```

You do not need to have information about every gene. In the example, nothing is known about Gene2, so the line after its name is left blank. If you have a list of genes and text information about them in a spreadsheet formatted as two columns with one row per gene, simply save this file as a tab-delineated text file.

SGD

A third Master Gene Table format is “**SGD**”. This is the format used for the list of genes in the Saccharomyces Genome Database (SGD), and is generally only relevant for yeast. As yeast comes pre-loaded in GeneSpring, details about this format are unnecessary.

Mapped

The fourth and most sophisticated Master Gene Table format is “**Mapped**”. Again, this format has one line per gene, with several fields separated by tabs. The first field (systematic name) must be present; all other fields are optional. The fields are described below. When creating your Master Gene Table, these fields should be entered in the order listed here.

1. **Systematic Name**: The normal way of referring to this gene. This name must be unique. The name entered in this field can be utilized by the Find Gene command to find this particular gene within GeneSpring. It is recommend that the name used as the gene’s systematic name be the name which labels that gene’s raw control strength values in your experiment data files. Any of this information can be accessed when you use the Find Gene command.
2. **Common Name**: An alternative way of referring to this gene. The name entered in this field can be utilized by the Find Gene command to find this particular gene within GeneSpring. Genes are not required to have a common name, and common names do not have to be unique, although duplicated common names may lead to confusion if the common name is how the gene is referred to in the experiment files. This information can be accessed when you use the Find Gene command.
3. **Map**: Mapping information for this gene. Sequence position, for example, a first chromosome gene would be 1:228836..229309 inclusive. For an example of the mapped Cytogenetic position (such as 16q12.1).
4. **EC number**: The EC number for this gene, if known.

Copyright 1998-2001 Silicon Genetics

5. **Description**: A description of this gene, if known. This information can be accessed when you use the Find Gene command.
6. **Product**: The protein product coded for by this gene, if known. This information can be accessed when you use the Find Gene command.
7. **Phenotype**: A description of the phenotype for this gene, if known.
8. **Function**: A description of the function of this gene product, if known.
9. **Keywords**: Keywords associated with this gene, if known. Separate keywords with semicolons. This information can be accessed when you use the Find Gene command.
10. **GenBank Accession Number**: The GenBank identifier for this gene, if known. If the GenBank identifiers for your genes were not used as either their systematic or common names, then including the GenBank Accession Number in this field allows you to update the information about this particular gene directly from GenBank. See "Updating your Master Gene Table with GeneSpider" on page 2-15 for more information.
11. **Synonym**: This column allows for other names to be entered for the genes. Multiple names should be separated by semicolons (;).
12. **Sequence**: The sequence data, if known.
13. **PM**: The Public Medline accession number, if known. Multiple identifiers should be separated by semicolons (;).
14. **custom1**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.
15. **custom2**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.
16. **custom3**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.
17. **Type**: A result of the conversion from a *.gbk* file to a master table of genes. It come from the GenBank column "feature type". For example, possible entries include: CDS, gene, terminator, rRNA.
18. **Database reference (also called DBid)**: A specific field returned by the GeneSpider. There are *dbxref* entries in GenBank, and these entries give database ID for other, non-GenBank databases, such as the SwissProt ID numbers. There may be multiple entries for each gene.

Copyright 1998-2001 Silicon Genetics

The Mapped format allows you to link up to three different names (plus three more custom names) for the same gene. Using this method, you could query one gene using any of the data in the corresponding columns #A Systematic Name, #B Common Name, and #F Product. You can also describe genes in your overlay or do a search for a gene named in column #2 Common Name and find the corresponding accession number.

The titles are included here only for clarity. Remember, when you are using the "mapped" format, you must include any blank fields in their appropriate columns. The gene's systematic name should always be in the first column, its common name in the second, and its mapping information in the third column, and so on, even if the second column is completely blank because there are no common names for any of your genes.

GenBank or EMBL Files

If you use a single GenBank file to describe the genome, you do not have to use a Master Gene Table and therefore do not have to enter any of the information discussed in "What Format do these Data Need to be in?" on page -1. Nor do you need a separate file to contain the sequence data (the files for sequence data are described in "Sequence Data" on page -5). The GenBank file can be downloaded directly from GenBank, if you open a web browser to the URL of the organism you are installing. For example, "ecoli.gbk" is a 9.5-MB file, from the URL:

```
ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/Ecoli/
```

Generally this URL is the same for all of GenBank's bacterial genomes, with the name of the organism you are installing in place of "Ecoli". This URL may contain many file formats. Make certain to download the file with the suffix *.gbk*. An EMBL file may be used in place of a GenBank file.

Adding Extra Genes to a Genome Defined by a GenBank or EMBL file

You can use a GenBank or EMBL file to describe a genome and add in some extra genes. This is typically done to represent a strain slightly different from the sequenced strain. To do this you need to create a separate Master Gene Table containing all of the extra genes you wish to add. This file should be formatted using one of the four table of genes formats discussed in "What Format do these Data Need to be in?" on page -1.

If you are using an original *.gbk* file, you can simply go to their web site and update the entire file. Make sure you save it with the same name and to the same place as your current *.gbk* file.

Copyright 1998-2001 Silicon Genetics

To update GenBank information

1. In GeneSpring, open the genome you wish to update.

   a. Go to **`File > New Genome or Array`**. Another menu appears. The genomes included in this submenu depend on what genomes have been loaded into your copy of GeneSpring.

   b. Select the name of the genome you wish to update.

2. Go to **`Tools > GeneSpider > Update genes from GenBank`**.
3. Click the arrow to the right of the box labeled *What the spider will use to mine GenBank*. A drop-down menu will appear.
4. Click the name of the column in the table of genes containing the GenBank Accession Numbers.
5. Click the **`Start`** button. The GeneSpider will process GenBank's data, displaying how far it has gotten in the box labeled *Status*.
   If you get a dialog box with an error you can click the close button on the upper right hand corner of the error messages and continue the operation.
6. Type the name of the text file you would like the new Master Gene Table saved as in the box labeled *Save gene list to*. If you save the new Master Gene Table using the same name as the current table file (in this example, ORF_table.txt) then the updated file will define this genome, rather than the previous table of genes file. If you save this updated Master Gene Table under a different file name (for example, ORF_table2.txt), then the old Master Gene Table will continue to define the genome, although the updated Master Gene Table will have been saved in the same directory as the original Master Gene Table.
7. Click the **`Save and Close`** button to save the updated Master Gene Table. If, for some reason, you do not want to save, close the window by clicking the close button the upper right hand corner. You can select the **`Save and Close`** button at any time during the update. The searched items will have been temporarily stored in your computer and will be visible in GeneSpring when you restart. It will go through the genes it has already updated really fast. It will take five to 30 seconds per gene depending on how much data the GeneSpider is bringing back. You may want to let this program run over your lunch hour, or for very large genomes, overnight.

Sequence Data

GeneSpring loads in sequence data from a GenBank or EMBL file automatically. If you have sequence data that is not in a GenBank/EMBL file, then the sequence data should be put into a separate file and formatted using the *.seq* format. A severely abridged example of the *yeast.seq* file might look like the following.

```
>CHR1 Chromosome I data:
CCACACCACACCCACACACCCACACACCACCACCACACCACACCCACACACACA . . .
GTGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGGG
>CHR2 Complete DNA sequence of yeast chromosome II.
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGA . . .
```


Copyright 1998-2001 Silicon Genetics

```
AGAATAGGGTACTGTTAGGATTGTGTTAGGGTGTGGGTGTGGTGTGTGTGGG
TGTGGTGTGTGGGTGTGT
>CHR3 LOCUS       SCCHRIII   315341 bp    DNA             PLN
25-NOV-1996
CCCACACACCACACCCACACCACACCCACACACCACACACACCACACCCA . . .
AGTGTGTGGGTGTGGGTGTGTGGGTGTGGTGTGTGGGTGTGGTGTGTGTGGTGT
GTGGGTGTGGGTGTGTGGGTGTGGTGGGTGTGGTGTGTGTG
```

If you have multiple chromosomes, they should be named sequentially, CHR1, CHR2 and so on. If there is only one chromosome, name it CHR1.

The *.seq* format is *not* the same thing as the FASTA format. There is an example of the FASTA format at http://www.ncbi.nlm.nih.gov/BLAST/fasta.html.

Where Do I Put My Data Files?

The files should be put in the same folder within GeneSpring's data directory. The default data directory for GeneSpring in a PC is C:\Program Files\Silicon Genetics\GeneSpring\data. In this data directory, use your file management program to create a new sub-directory to hold the new genome data. This folder is usually named after the organism you are adding, but any memorable name will suffice.

There are three possible raw data files you may have when you create a new genome.

1. You must have a Master Gene Table or a GenBank/EMBL file(s).
2. You can have sequence data in *.seq* format.
3. You may have a file containing extra, non-GenBank genes (if you have any). The file of extra genes should be in one of the four standard Master Gene Table formats.

The three raw data files should all be placed within your new subdirectory.

Copyright 1998-2001 Silicon Genetics

Appendix I Installing a Genome from a Text File

The following steps are needed to load a genome. These steps are essentially the same as the questions you answer in the Genome Wizard. The specific examples and instructions given are for E. coli.

1. Open the GeneSpring data directory (typically C:\Program Files\SiliconGenetics\GeneSpring\data), using your file management program.
2. Create a sub-directory to hold the new genome data.
3. Copy your Master Gene Table, GenBank, or EMBL file(s) in this new directory. If you have a separate sequence file, put that in this new directory also. If you have a file containing extra genes, put that file in this new directory.
4. In the same directory, create a file describing the genome. The file name should end with *.genomedef*, such as *Ecoli.genomedef*. See "The .genomedef File" on page I-1, for what this file should contain.
5. All files within the "GeneSpring\data" directory (except those in the "cache" directory if there is one) ending in *.genomedef* are found automatically. Start GeneSpring to make sure your genome is properly loaded. You should be able to find its name by selecting **`File > New Genome`**. In this example "E. Coli" appears there.

Creating Folders for New Genomes

To manually create a new folder in the genome browser, you must go through a file management system, such as Windows Explorer®.

Before your new folder will appear in the navigator you will need to create a correct .genomedef file for that organisms. A *.genomedef* file will contain all the information GeneSpring needs to create a folder and other data objects. Make sure you save the *.genomedef* file in your new directory after you create it.

The *.genomedef* File

The *.genomedef* file contains a brief description of the genome. This file contains several lines, each of the form object-name space-colon-space object-value. For example: Object-name : object-value.

An example of how this actually appears in the *.genomedef* file is:

```
name : e.coli
```

In this example "name" is the object-name and "e.coli" is the object-value. The object-value can be thought of as the answer to the question posed by the object-name. In the *.genomedef* file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Blank lines and lines beginning with the number character (#) are ignored.

 Copyright 1998-2001 Silicon Genetics

Define Your Genome

This section is designed to help you create a *.genomedef* file for a particular genome, and therefore it is written as a series of questions for you to answer. There are two examples following each question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (bold-faced) example is an example of an actual answer to the question. Some of the lines the questions represent are required, and others are not, each question will be annotated accordingly. The genome, "e.coli" is used as the example throughout this section.

1. Enter the name of your genome as you wish it to appear in GeneSpring. This line is required.

```
name : the name of the genome
```

```
name : e.coli
```

2. If you are using a Master Gene Table to define your genome, enter the complete file name of the file containing the Master Gene Table. This question and the next question are mutually exclusive, you must have one of them in your *.genomedef* file.

```
ORFs : the complete file name of the file containing the Master Gene Table of all the genes
```

```
ORFs : genelist.txt
```

3. If you are using either a GenBank file or an EMBL file to define your genome, enter the complete file name of the file describing your genome. This is necessary if you used a GenBank or EMBL file. This question and the previous question are mutually exclusive. One of the two is required.

```
GenBank: the name of the GenBank/EMBL file describing this genome
```

```
GenBank : ecoli.gbk
```

   Or,

```
GenBank : ecoli.ebl
```

Even if you are using an EMBL file the object-name in this entry is GenBank.

4. If you have a file containing extra genes, enter the complete file name of the file containing these supplementary elements. This line is optional, but must be included in the *.genomedef* file for GeneSpring to incorporate this data.

```
nonORFs : the complete file name of the extra file containing other genomic elements than in the ORFs file
```

```
nonORFs : extragenes.txt
```

Copyright 1998-2001 Silicon Genetics

5. If you have a file containing the sequence data for the genome, enter the complete name of that file, including the *.seq* suffix. This line is optional, but must be included in the *.genomedef* file for GeneSpring to incorporate sequence data not included within a GenBank or EMBL file.

```
sequence : the name of a file containing the
sequence(s) for the genome

sequence : ecoli.seq
```

6. If you are using a Master Gene Table to define your genome, indicate which format you used. The four Master Gene Table format options are: name list, name function, SGD, or mapped. These are also the four possible object-values for this question. See "What Format do these Data Need to be in?" on page H-1 for a description of these formats. This line is required if the ORFs line from question two was used.

```
ORFFormat : the format for the Master Gene Table
specified in the ORFs line

ORFFormat : mapped
```

7. If you are using a supplementary table of genes file, indicate which table of genes format is used in this file. This can be one of the four table of genes format options: name list, name function, SGD, or mapped. These are also the four possible object-values for this question. See "What Format do these Data Need to be in?" on page H-1 for a description of these formats. This line is required if the nonORFs line from question four was used and the format for this file is different from the format given in response to question six.

```
nonORFFormat : the format for the file specified
in the nonORFs if different from the file of ORFs

nonORFFormat : name function
```

8. If the genome you are entering has been sequenced, then you should answer "true" to this question. This line is optional, but if you are using a GenBank file, an EMBL file, or a *.seq* file to define your organism's sequence, then the sequence data will not be loaded into GeneSpring if this line is not in the *.genomedef* file. If your organism has not been sequenced, or you do not have its sequence information available, then you do not need to enter this line in the *.genomedef* file.

```
KnowGenome : set to true if the genome is
sequenced, and false if not

KnowGenome : true
```

Copyright 1998-2001 Silicon Genetics

9. If the genome you are entering is a circular genome (such as bacteria, plasmids, and viruses) then you should answer "true" to this question. This line is optional, if you do not enter it, or answer it "false" then your genome will not be plotted as a circle in the physical position display.

```
CircularGenome : set to true if the genome should
be plotted as a circle and false otherwise

CircularGenome : true
```

10. Are there web-based databases you would like to be able to link to automatically? If not, skip this question. You can link to the URL of any web-based database containing the name of your gene. Each separate link should consist of one line in the *.genomedef* file. Each line should start with the phrase "GeneHypertextLinks" followed by a colon, followed by the description of the link.

    The description of the link is the name of the link (the name you want to appear on a button in GeneSpring), which *must* be followed by a colon, not a semicolon. Any field in angle brackets (for example, <field>) will be replaced by the value of that parameter. The allowed parameters are:

    - systematic
    - common
    - genbank
    - ec
    - pubmed
    - map
    - chromosome
    - synonyms
    - description
    - phenotype
    - function
    - product
    - keywords
    - dbid
    - custom1
    - custom2
    - custom3

    A link will only be enabled for a particular gene if *all* parameters mentioned in that URL are defined for that gene.

```
GeneHypertextLinks : Links to external web based
databases. You can have more than one of these
lines; you should have one line for each link.

GeneHypertextLinks : linkname:http://www.some-
where.org&gene=<systematic>&id=<genbank>
```


Copyright 1998-2001 Silicon Genetics

This example should be one consecutive line beginning “GeneHypertextLinks : ”, but is has been broken into separate lines to allow it to fit on this page. It should be entered into the file as single line, without carriage returns. There is no space between the semicolon following the link’s name and the associated URL.

Experiment URLs work exactly the same way, except that they begin with ExperimentHypertextLinks instead of GeneHypertextLinks, and the things in <> signs are the names of parameters. A link will only be shown in the Experiment Inspector if the experiment has parameters with names matching all fields in the URL.

In both cases, the parameter names are *not* case sensitive, so if an experiment has a parameter called Time, you can specify it as <time>, <Time>, or <TIME> in the URL, and they will all work.

```
ExperimentHypertextLinks : Links to external web
based databases. You can have more than one of
these lines; you should have one line for each
link.

ExperimentHypertextLinks : linkname:http://
www.somewhere.experimentlikemine=<system-
atic>&id=<time>
```

11. Use this line if there is a particular experiment you would like GeneSpring to automatically display in the genome browser when you open this genome. This *.genomedef* entry is optional, if it is not included GeneSpring will open the genome but not open any particular experiment when you select this genome to be displayed.

```
defaultExperiment : the name of the default exper-
iment you want started when opening this genome

defaultExperiment : yeast extraterrestrial studies
```

The name following the object-value should be the same name given to the experiment in the name line of its *.html* file and/or it should be the name entered for the experiment in the *Properties of an Experiment Set* panel of the New Experiment Wizard. Both of these options are case sensitive, so make sure the spelling and capitalization is correct. See “The Experiment Wizard” on page D-1 for more information about entering an experiment. If you do not know the name of any experiment done with this genome when you create it, this line can be added or modified afterwards. (Just remember to save the modified *.genomedef* file.)

 Copyright 1998-2001 Silicon Genetics

12. If you work in a group that is storing data and analyses in a shared environment (usually this means that you have all of the data for the group in one file system) you will probably also want to have your own local data for each genome. A specific use of this is for gene lists (not the genome defining Master Gene Table, but a gene list you create within GeneSpring): it is often desirable for each person to keep the gene lists they create initially separate as trial lists, and then merge them into the groups' permanent set when they are more certain about the significance of individual lists. To store data locally, you specify (in the *.genomedef* file of each genome) a second directory to be searched for experiment data, gene lists, trees, etc. This directory is specified with the line below. This is an optional line.

```
HomeDirectory : The complete path of an extra
directory to search for to find information for
this genome
```

```
HomeDirectory : C:\Silicon Genetics\Gene-
Spring\data\Ecoli
```

    Including this line means that both this directory on your local computer and the directory containing the *.genomedef* file are searched for experiment data, gene lists, classifications, and so forth. As the local directory must be indicated in the shared directory, every user in your group must keep their local directory in the same place on their local computers. In the example this place would be the C:Silicon Genetics\GeneSpring\data\Ecoli.

13. If there is a prefix (a string of characters) prepended to the start of your genes' systematic names you can tell GeneSpring to disregard this first part of the gene name and not display it. This line is not required, and it is rarely used.

```
SystematicPrefix : a string that is often
prepended to the start of gene names, and should
be ignored if seen
```

```
SystematicPrefix : ecoli/
```

14. If you wish the genes' systematic names to appear entirely in upper case letters, GeneSpring can convert them to this automatically. This line is not required, and is rarely used.

```
ForceUpperCase : set to true if you want all the
names of the genes converted to upper case, set
this line to false otherwise
```

```
ForceUpperCase : true
```

15. If you wish the genes' systematic names to appear entirely in lower case letters, GeneSpring can convert them to this automatically. This line is not required, and is rarely used.

```
ForceLowerCase : set to true if you want all the
names of the genes converted to lower case, set
this line to false otherwise
```

```
ForceLowerCase : false
```


Copyright 1998-2001 Silicon Genetics

16. You can place any data you wish in the custom label columns.

```
Custom1Label : heading
```

```
Custom1Label : interacts with P53
```

17. You can place any data you wish in the custom label columns.

```
Custom2Label : heading
```

```
Custom2Label : molecular weight
```

18. You can place any data you wish in the custom label columns.

```
Custom3Label : heading
```

```
Custom3Label : plate and well location
```

19. If your genome has a unique identifier, such as a nickname, that would speed searching for it, enter it in this line.

```
Identifier : optional unique identifier for the
whole genome
```

```
Identifier : dutch elm disease study
```

20. You can use ChromosomeNames to cause the “mito” chromosomes to be sorted separately from the remaining chromosomes.

```
ChromosomeNames :
```

```
ChromosomeNames : I;II;III;IV;V;VI;VII;VIII;IX;X;
XI;XII;XIII;XIV;XV;XVI;mito
```

21. You can set your genome to be able to find genes with the same names in other genomes. There are two ways to set up the *.genomedef* file, as shown below. For details on this feature, please refer to “Making Lists of Homologs and Orthologs” on page 4-31.

```
AcceptedDirectTranslation : [genome1];[genome2]
```

Or:

```
AcceptedDirectTranslation : [genome1]
AcceptedDirectTranslation : [genome2]
```

Make sure you save the *.genomedef* file after you create it.

 Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Appendix J Installing from a Text File

This is possibly the most tedious and unforgiving of the experiment loading methods. However, it is necessary to be at least slightly familiar with the methods, as you will need to change the *experiment* file (or re-enter your experiment through another method) when you need to make changes to the experiment.

Generally, an *.experiment* file is a text file describing where the data file(s) are, what their format is, what the parameters for the experiment are, and what normalizations need to be done. You can also specify pictures to be associated with the files, and various other things. Each line in an *.experiment* file is either blank or a line of the form object-name space-colon-space object-value:

```
Object-name : object-value
```

An example of this is:

```
name : Yeast extraterrestrial studies
```

Obviously, "name" is the object-name and "Yeast extraterrestrial studies" is the object-value. The object value can be thought of as the answer to the question posed by the object-name. In the *.experiment* file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Usually, when an experiment looks like it is not installed correctly it is because of a spelling or capitalization error.

Due to the complexity of the information contained in the *.experiment* file, this section is designed to help you create a *.experiment* file for a particular experiment, rather than explaining exactly what each possible answer means. There are two examples following each question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (bold-faced) example is an example of an actual answer to the question. A fictitious experiment, "Yeast extraterrestrial studies", is used as the example experiment throughout this chapter. A complete *.experiment* file for the "Yeast extraterrestrial studies" experiment is given in this chapter. There are eighteen sections and thirty eight questions which must be answered in their presented order.

Define Your Experiment

1. Enter the name of your experiment or samples as you wish it to appear in the GeneSpring menu system.

   ```
   name : Your experiment name here
   ```

   `name : Yeast extraterrestrial studies`

2. How many samples are there in the experiment you have just named? A sample is defined as each time a numerical measurement is taken for your entire set of genes.

   ```
   Experiments : The number of samples
   ```

   `Experiments : 40`

Copyright 1998-2001 Silicon Genetics

3. How many different parameters were taken? A parameter is used to describe the condition (or conditions) in the experiment. See “Definitions of Parameters” on page 2-11 for a more through description of parameters.

```
Parameters : The number of parameters

Parameters : 4
```

4. Name the parameters:

```
Parameter#Name : Name of the indicated parameter
```

   Make sure to name each of the parameters enumerated in question 3.

```
Parameter1Name : Kryptonite concentration
Parameter2Name : Variety of yeast
Parameter3Name : Test repeat number
Parameter4Name : Andromeda Strain infection
```

Define Your Parameters

In number 4 of section “Define Your Experiment” on page J-1 you named and numbered each parameter. They will be referred to by their number for the remainder of this example. For reasons of brevity, the questions in this section are all phrased in reference to parameter 1, but you should answer each question for every parameter enumerated in question 4.

5. If there are units associated with parameter 1, name them.

```
Parameter#Units : name of the units associated
with the indicated parameter
```

   If a parameter does not have a unit name associated with it, either do not enter the line “Parameter#Units : ” for the parameter without units, or enter the object-name “Parameter#Units” and the space-colon-space, but leave the name of the units (the object-value) blank.

```
Parameter1Units : ppm
Parameter2Units :
Parameter3Units :
Parameter4Units :
```

6. Is parameter 1 defined by a number, i.e. are the parameter values associated with parameter 1 numbers? If the answer is yes, enter “true” after “Parameter1IsNumber : ” and if the answer is no, enter “false”.

```
Parameter#IsNumber : enter either true or false

Parameter1IsNumber : true
Parameter2IsNumber : false
Parameter3IsNumber : true
Parameter4IsNumber : false
```

Copyright 1998-2001 Silicon Genetics

7. This question is only applicable to those parameters defined by a number. (I.e. for those parameters for whom the answer to question 6 is true.) Would you like the number defining parameter 1 graphed on a logarithmic scale? If this answer is yes, enter "true" as the object-value following "Parameter1IsLogarithmic". If the answer is no, either do not enter the "Parameter1IsLogarithmic : " line, or type "false" as the object-value. The answer to this question is automatically false if a number does not define the parameter in question.

```
Parameter#IsLogarithmic : enter either true or
false

Parameter1IsLogarithmic : false
Parameter2IsLogarithmic : false
Parameter3IsLogarithmic : false
Parameter4IsLogarithmic : false
```

8. Of the following four choices, choose the most appropriate display for parameter 1. (You may alter your choice within GeneSpring, the display you are indicating here will simply be the default display). See "Definitions of Parameters" on page 2-11 for more details about each of these display options.

   - Parameter 1 is continuous. This means when you are graphing the data by this parameter the data points will be connected together by lines instead of being graphed as discrete points. Follow "Parameter1IsContinuious" with true if this is how you wish the parameter to be graphed. If one of the other possibilities seems more correct for parameter 1, either enter "false" as the object-value, or do not include the line beginning with "Parameter1IsContinuious".

```
Parameter#IsContinuous : either true or false

Parameter1IsContinuous : true
Parameter2IsContinuous : false
Parameter3IsContinuous : false
Parameter4IsContinuous : false
```

   - Parameter 1 is a category (or set of categories) and you wish to color code the display by their membership. If this is the display you wish for parameter 1, answer the object-name lines, "Parameter1IsContinuous", "Parameter1IsSet", and "Parameter1IsRepeat"all with the object-value "false".

```
This is the case for parameter 2 in the Yeast can-
cer time series experiment.
```

Copyright 1998-2001 Silicon Genetics

- Parameter 1 is a replicate parameter by which you do not wish to distinguish information graphically. Follow "Parameter1IsRepeat" with the object-value "true" if this is how you wish this parameter to be graphed. If one of the other possible parameters interpretations is correct for parameter 1, either enter "false" as the object-value, or do not include the line beginning with "Parameter1IsRepeat".

```
Parameter#IsRepeat : either true or false

Parameter1IsRepeat : false
Parameter2IsRepeat : false
Parameter3IsRepeat : true
Parameter4IsRepeat : false
```

- You wish to use parameter 1 to separate the data into discrete graphs viewed next to each other on the same screen. This is a non-continuous parameter. Follow "Parameter1IsSet" with the object-value "true" if this is how you wish this parameter to be displayed. If one of the other possibilities seems more correct for parameter 1, either enter "false" as the object-value, or do not include the line beginning with "Parameter1IsSet".

```
Parameter#IsSet : either true or false

Parameter1IsSet : false
Parameter2IsSet : false
Parameter3IsSet : false
Parameter4IsSet : true
```

9. Enter the number or label applicable to each sample, as it is associated with parameter 1. This is where you tell GeneSpring what each condition means, as far as each parameter is concerned.

```
Parameter#Experiment# : either a value or a name
associated with both the parameter indicated and
the sample indicated.
```

For each parameter you must indicate a label to associate with every condition.

```
Parameter1Experiment1 : 0
Parameter1Experiment2 : 10
Parameter1Experiment3 : 20
Parameter1Experiment4 : 30
Parameter1Experiment5 : 40
Parameter1Experiment6 : 0
Parameter1Experiment7 : 10
Parameter1Experiment8 : 20
Parameter1Experiment9 : 30
Parameter1Experiment10 : 40
Parameter1Experiment11 : 0
Parameter1Experiment12 : 10
. . .
Parameter2Experiment1 : A
Parameter2Experiment2 : A
```


Copyright 1998-2001 Silicon Genetics

```
Parameter2Experiment3 : A
Parameter2Experiment4 : A
Parameter2Experiment5 : A
Parameter2Experiment6 : B
Parameter2Experiment7 : B
Parameter2Experiment8 : B
Parameter2Experiment9 : B
Parameter2Experiment10 : B
Parameter2Experiment11 : A
Parameter2Experiment12 : A
. . .
Parameter3Experiment1 : Test 1
Parameter3Experiment2 : Test 1
Parameter3Experiment3 : Test 1
Parameter3Experiment4 : Test 1
Parameter3Experiment5 : Test 1
Parameter3Experiment6 : Test 1
Parameter3Experiment7 : Test 1
Parameter3Experiment8 : Test 1
Parameter3Experiment9 : Test 1
Parameter3Experiment10 : Test 1
Parameter3Experiment11 : Test 1
Parameter3Experiment12 : Test 1
Parameter3Experiment13 : Test 1
Parameter3Experiment14 : Test 1
Parameter3Experiment15 : Test 1
Parameter3Experiment16 : Test 1
Parameter3Experiment17 : Test 1
Parameter3Experiment18 : Test 1
Parameter3Experiment19 : Test 1
Parameter3Experiment20 : Test 1
Parameter3Experiment21 : Test 2
Parameter3Experiment22 : Test 2
Parameter3Experiment23 : Test 2
Parameter3Experiment24 : Test 2
. . .
Parameter4Experiment1 : healthy
Parameter4Experiment2 : healthy
Parameter4Experiment3 : healthy
Parameter4Experiment4 : healthy
Parameter4Experiment5 : healthy
Parameter4Experiment6 : healthy
Parameter4Experiment7 : healthy
Parameter4Experiment8 : healthy
Parameter4Experiment9 : healthy
Parameter4Experiment10 : healthy
```


Copyright 1998-2001 Silicon Genetics

```
Parameter4Experiment11 : Andromeda strain
Parameter4Experiment12 : Andromeda strain
Parameter4Experiment13 : Andromeda strain
. . .
```

In order to illustrate how to write all four of the possible parameter displays, the Yeast extraterrestrial study is a fairly large experiment, with many samples, as well as many parameters. This makes the entry for question 9 extremely long. You may well have a much smaller and less complex set of notations to write down.

Describe your Data Files

10. Are all of your samples in the same data file? If so enter this:

```
DataFileName : complete name of the file containing your experimental data
```

```
DataFileName : array.txt
```

If even one of your experiment's samples are in a separate file from the rest, you must specify a separate file name for each sample.

```
Experiment#FileName : complete name of the file containing the data from the sample indicated
```

```
Experiment1FileName : 1A0.txt
Experiment2FileName : 1A10.txt
Experiment3FileName : 1A20.txt
Experiment4FileName : 1A30.txt
Experiment5FileName : 1A40.txt
Experimetn6FileName : 1B0.txt
Experiment7FileName : 1B10.txt
Experiment8FileName : 1B20.txt
Experiment9FileName : 1B30.txt
Experiment10FileName : 1B40.txt
Experiment11FileName : 1AndromedaA0.txt
Experiment12FileName : 1AndromedaA10.txt
Experiment13FileName : 1AndromedaA20.txt
Experiment14FileName : 1AndromedaA30.txt
Experiment15FileName : 1AndromedaA40.txt
Experimetn16FileName : 1AndromedaB0.txt
Experiment17FileName : 1AndromedaB10.txt
Experiment18FileName : 1AndromedaB20.txt
Experiment19FileName : 1AndromedaB30.txt
Experiment20FileName : 1AndromedaB40.txt
Experiment21FileName : 2A0.txt
Experiment22FileName : 2A10.txt
Experiment23FileName : 2A20.txt
```


Copyright 1998-2001 Silicon Genetics

```
Experiment24FileName : 2A30.txt
Experiment25FileName : 2A40.txt
Experimetn26FileName : 2B0.txt
Experiment27FileName : 2B10.txt
Experiment28FileName : 2B20.txt
Experiment29FileName : 2B30.txt
Experiment30FileName : 2B40.txt
Experiment31FileName : 2AndromedaA0.txt
Experiment32FileName : 2AndromedaA10.txt
Experiment33FileName : 2AndromedaA20.txt
Experiment34FileName : 2AndromedaA30.txt
Experiment35FileName : 2AndromedaA40.txt
Experimetn36FileName : 2AndromedaB0.txt
Experiment37FileName : 2AndromedaB10.txt
Experiment38FileName : 2AndromedaB20.txt
Experiment39FileName : 2AndromedaB30.txt
Experiment40FileName : 2AndromedaB40.txt
```

Data File Header Lines

If you have more than one data file, and they have different column layouts, then you must answer these questions for *every* experiment/sample data file you have.

11. Does your data file have one or more headlines not containing experimental data?

```
Headlines : number of headlines in the data file

Headlines : 1
```

If your data files all use different layouts, but all of them have the same number of headlines, you may use the general object-name given above, rather than entering the number of headlines for each data file. If you have more than one data file, with different numbers of headlines use the object-name given below. If you are doing this, make sure to indicate the number of headlines for every sample.

```
Experiment#Headlines : number of headlines in the
data file of the experiment indicated

Experiment1Headlines : 1
Experiment2Headlines : 3
Experimetn3Headlines : 1
. . .
```


Copyright 1998-2001 Silicon Genetics

Gene Names

12. Which column of your data file contains the gene name?

```
GeneColumn : number of the column the gene name is
found in

GeneColumn : 1
```

If your data files all have a different column layout, but all of them have the gene name in the same column, you may use the general object-name given above, rather than entering the column number of the gene name for each data file. If you have more than one data file with different column layouts, and they have different columns containing the gene name, use the object-name given below. If you are doing this, make sure to indicate the column containing the gene name for every sample.

```
Experiment#GeneColumn : number of the column the
gene name is found in, for the experiment indi-
cated

Experiment1GeneColumn : 2
Experiment2GeneColumn : 3
Experiment3GeneColumn : 2
. . .
```

Explain to GeneSpring how to locate only the Gene Name

These questions are only applicable if the column containing the gene name contains other notations as well, notations not occurring in the list of genes defining the genome. If column containing the gene names in your data file(s) only contains the gene name as it appears in the table of genes file or the GenBank/EMBL file defining this genome, skip these two questions and do not enter the lines associated with them in your *.experiment* file.

13. GeneSpring can remove a set suffix from a gene name. A set suffix is a fixed string of characters which appear frequently at the end of your genes.

```
RemoveGeneSuffix : exact suffix you wish removed
from the gene name

RemoveGeneSuffix : _at
```

14. GeneSpring can remove the entire notation following a slash (/), including the slash itself. To do this, enter “true” as the object-value. To ignore this ability, thus leaving the gene name alone either enter “false” as the object-value after “RemoveSlash : ” or do not include this line in your *.experiment* file.

```
RemoveSlash : either true or false

RemoveSlash : true
```


Copyright 1998-2001 Silicon Genetics

Explain to GeneSpring How to Read the Region Specifications

Skip these questions, and their associated entries in the *.experiment* file, if the samples in your experiment did not involve multiple arrays or sections of arrays needing to be normalized separately.

15. If your experiment used multiple arrays, or sections of arrays, needing to be normalized separately, indicate to GeneSpring which column of your data file indicates the region of the array, and/or which array a particular gene reading came from.

```
RegionColumn : number of the column the region
specification is found in

RegionColumn : 1
```

If your data files all have a different column layout, but all of them have the region specification in the same column, you may use the general object-name given above, rather than entering the column number of the region specification for each data file. If you have more than one data file with different column layouts, and they have different columns containing the region specification, use the object-name given below. If you are doing this, make sure to indicate the column containing the region specification for every sample.

```
Experiment#RegionColumn : number of the column the
region specification is found in, for the experi-
ment indicated

Experiment1RegionColumn : 1
Experiment2RegionColumn : 2
Experiment3RegionColumn : 1
. . .
```

The required *.layout* file for Region Specifications

16. If you have region specifications you must have a *layout* file. (See "The Layout file" on page K-2 for everything this file can or should contain.) Tell GeneSpring where to find this file:

```
Layout : complete name of the layout file

Layout : AffyYeastLayout4.txt
```

Locate the Data Column

17. Which column of your data file contains the raw data reading for Sample 1?

```
Experiment#IntensityColumn : number of the column
containing the raw data for the sample indicated
```


Copyright 1998-2001 Silicon Genetics

```
Experiment1IntensityColumn : 4
Experiment2IntensityColumn : 9
Experiment3IntensityColumn : 14
Experiment4IntensityColumn : 19
Experiment5IntensityColumn : 24
Experiment6IntensityColumn : 29
Experiment7IntensityColumn : 34
. . .
```

If your data is all in the same file you will have to indicate the raw data column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the raw data. On the other hand, if you have separate data files, with the same column containing the raw data you may use the general object-name given below, rather than entering the column number of the raw data for each file.

```
IntensityColumn : number of the column containing
the signal intensity data

IntensityColumn : 7
```

18. If your data file has a column indicating the background signal, tell GeneSpring which column contains that information. If your data does not have a background reading, skip this question, and the associated *.experiment* file entry.

```
Experiment#IntensityBackColumn : number of the
column containing the background reading for the
sample indicated

Experiment1IntensityBackColumn : 5
Experiment2IntensityBackColumn : 10
Experiment3IntensityBackColumn : 15
Experiment4IntensityBackColumn : 20
Experiment5IntensityBackColumn : 25
Experiment6IntensityBackColumn : 30
Experiment7IntensityBackColumn : 35
. . .
```

If your data is all in the same file you will have to indicate the background reading column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the background data. If, on the other hand, you have separate data files, with the same column containing the background data you may use the general object-name given below, rather than entering the column number of the background data for each file.

```
IntensityBackColumn : number of the column con-
taining the background reading

IntensityBackColumn : 8
```


Copyright 1998-2001 Silicon Genetics

The Control Channel Value

These questions only apply if your sample has a control channel, which is generally only applicable to two-color experiments, such as Incyte or Sentini experiments. If your data does not have control channel values, skip this section and the associated *.experiment* file entries.

19. If your data has control channel values, which column of your data file gives the reference value? If your data does not have control channel values, skip this question, and the associated *.experiment* file entry.

```
Experiment#ReferenceColumn : number of the column
containing the control channel values for the
experiment indicated

Experiment1ReferenceColumn : 6
Experiment2ReferenceColumn : 11
Experiment3ReferenceColumn : 16
Experiment4ReferenceColumn : 21
Experiment5ReferenceColumn : 26
Experiment6ReferenceColumn : 31
Experiment7ReferenceColumn : 36
. . .
```

If your data is all in the same file you will have to indicate the reference column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the control channel values. On the other hand, if you have separate data files with the same column containing the control channel values, you may use the general object-name given below, rather than entering the column number for the control channel values in each file.

```
ReferenceColumn : number of the column containing
the control channel values

ReferenceColumn : 9
```

20. If your data includes the control channel's background signal, which column of your data file contains that information? If your data does not have control channel values, skip this question, and the associated *.experiment* file entry.

```
Experiment#ReferenceBackColumn : number of the
column containing the control channel's background
signals for the sample indicated

Experiment1ReferenceBackColumn : 7
Experiment2ReferenceBackColumn : 12
Experiment3ReferenceBackColumn : 17
Experiment4ReferenceBackColumn : 22
Experiment5ReferenceBackColumn : 27
Experiment6ReferenceBackColumn : 32
Experiment7ReferenceBackColumn : 37
. . .
```


Copyright 1998-2001 Silicon Genetics

If your data is all in the same file you will have to indicate the control channel background column for each experiment, illustrated above. This is also true if you have two or more data files with different columns containing the control channel's background values. If, on the other hand, you have separate data files, with the same column containing the control channel's background values, you may use the general object-name given below, rather than entering the column number of the control channel's background values for each file.

```
ReferenceBackColumn : number of the column con-
taining the control channel's background values

ReferenceBackColumn : 10
```

Measurement Flags

21. If your data file has a notation (flag) indicating whether or not the experiment worked for each gene, indicate which column contains this information. If your data does not include this information, skip this question, and the associated *.experiment* file entries.

```
Experiment#OkColumn : number of the column saying
whether or not the experiment indicated worked for
each gene

Experiment1OkColumn : 8
Experiment2OkColumn : 13
Experiment3OkColumn : 18
Experiment4OkColumn : 23
Experiment5OkColumn : 28
Experiment6OkColumn : 33
Experiment7OkColumn : 38
. . .
```

If your data is all in the same file you will have to indicate the *experiment worked* column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the *experiment worked* information. If, on the other hand, you have separate data files, with the same column containing the *experiment worked* notation, you may use the general object-name given below, rather than entering the column number of the reference's background values for each file.

```
OkColumn : number of the column saying whether or
not the experiment worked for each gene

OkColumn : 11
```


Copyright 1998-2001 Silicon Genetics

22. If you have a column indicating whether or not your experiment worked, what is the designation used in this column to indicate the experiment worked? (Often this is just a letter, such as P for Present or Passed.) If you do not have an *experiment worked* column, skip this question and the associated *.experiment* entry.

```
StatusOkString : the value, letter or word indi-
cating the sample is ok to use
```

```
StatusOkString : P
```

You can have more than one entry indicating the status. If you were not sure if your experiment recorded P for passed or O for OK, place both in the line, separated by vertical bars.

You might also have a designation for Marginal or Questionable data. (Often this is just a letter, such as M for Marginal.)

```
StatusMarginalString : the value, letter or word
indicating the sample is of marginal quality
```

```
StatusMarginalString : M|Q
```

You might also have a designation for Failed or Absent data. (Often this is just a letter, such as A for Absent.)

```
StatusFailedString : the value, letter or word
indicating the sample is absent
```

```
StatusFailedString : F|A
```

Associating a Picture with a Sample

Pictures are nice, but they are not necessary. If you don't have any, skip this section and the associated *.experiment* file entries.

23. If you have any pictures you wish to associate with any or all of the samples use the line given below to tell GeneSpring where to find the picture. If you do not have a picture to associate with every sample, GeneSpring will display the picture associated with the next closest sample with an associated picture.

```
Experiment#Image : the complete file name of the
file containing the picture to associate with the
indicated file
```


Copyright 1998-2001 Silicon Genetics

If you have a picture associated with every sample this section of your *.experiment* file should look similar to this:

```
Experiment1Image : yeastpict1A0.gif
Experiment2Image : yeastpict1A10.gif
Experiment3Image : yeastpict1A20.gif
Experiment4Image : yeastpict1A30.gif
Experiment5Image : yeastpict1A40.gif
Experiment6Image : yeastpict1B0.gif
Experiment7Image : yeastpict1B10.gif
. . .
```

If you have only one picture to associate with the entire experiment being described in your *.experiment* file, the picture entry should look similar to this one:

```
Experiment1Image : happy_yeast_picture.gif
```

If you have some pictures to associate with some but not all points in your sample the picture entries in your *.experiment* file should look similar to these:

```
Experiment1Image : yeastpict1A.gif
Experiment6Image : yeastpict1B.gif
Experiment11Image : yeastpict1AndromedaA.gif
Experiment16Image : yeastpict1AndromedaB.gif
Experiment21Image : yeastpict2A.gif
Experiment26Image : yeastpict2B.gif
Experiment31Image : yeastpict2AndromedaA.gif
Experiment36Image : yeastpict2AndromedaB.gif
```

Normalizations: Negative Controls

24. Do you have any genes designated as negative controls on your array? You have negative controls when there is DNA from a different genome than the one you are investigating on the array. Entering "true" as the object-value of the line given below means you have negative controls, and you want GeneSpring to normalize your samples using the negative control values. This normalization method takes the average signal intensities for all of the negative controls and subtracts this number from the signal intensity of each gene. For more info about this normalization option, see "Normalizing Options" on page G-1. If you do not have negative controls, or do not want to normalize your samples using the data from them, either do not enter the "NormalizeNegControl : " line, or type "false" as the object-value.

```
NormalizeNegControl : either true or false
```

```
NormalizeNegControl : false
```

 Copyright 1998-2001 Silicon Genetics

The required layout file for negative controls

25. If you do not have negative controls or are not using them to normalize your data, skip this question and the associated *.experiment* file entry. If you are using negative controls you must have a layout file. (See "The Layout file" on page K-2 for what this file can or should contain.) There are two normalization options requiring you to have a layout file. They both use this line to tell GeneSpring where to find the layout file. You should only have one layout file, and you should only enter the line, "Layout : name of layout file", once. You may have entered this file already, please refer to "The required .layout file for Region Specifications" on page J-9.

```
Layout : complete name of the layout file
```

```
Layout : AffyYeastLayout4.txt
```

Normalizations: Control Channel Values

If you do not have control channel values, skip these questions and the associated *.experiment* file entries.

26. If you have a control channel value for each gene to indicate the trust you have in the experimental data for each gene you probably want to normalize the genes by dividing their control strength by the control channel's control strength. If you have a background signal for either or both of these values, it is subtracted from the signal intensities before they are divided. For more information on this normalization option, see "Normalizing Options" on page G-1. If you wish to use this normalization, enter "true" as the object-value in the line illustrated below. If you do not have control channel values, or you do not wish your data to be normalized using the control channel values, either do not enter the line "NormalizeToReference : ", or enter "false" as the object-value in that line. Control channels generally apply to two-color experiments.

```
NormalizeToReference : either true or false
```

```
NormalizeToReference : true
```

27. If you do not have control channel values, skip this question and the associated *.experiment* file entry. Sometimes the control channel value is very low and would artificially inflate the noise for its gene, indicate the minimum value you would be willing to divide a gene's signal by:

```
NormalizeMinControl : the minimum signal value to
be used as a reference value for normalization
purposes
```

```
NormalizeMinControl : 10
```

    If you do not enter this line in your *.experiment* file and you do have control channel values, GeneSpring will automatically use the value given here, 10, as the default cut-off value.

 Copyright 1998-2001 Silicon Genetics

28. If you have control channel values for your experiment, but the column containing the "raw data" has already been normalized using this information (for example, your data is reported in ratio form), you can tell GeneSpring this, using the line illustrated below. If you have the raw data from both the gene and its control it is suggested you let GeneSpring perform your normalization, rather than using this option. For example, Incyte data is reported in what they call "ratio" form, but the ratio reported is not actually the gene's signal divided by its control; in this case it would probably be better to use the raw signal and control values and let GeneSpring perform the normalization. If you want to go ahead and use previously normalized data as your raw data, you should still tell GeneSpring in which column(s) the control signals are located.

```
UseReferenceAsStrength : enter true or false
```

```
UseReferenceAsStrength : false
```

Normalizations: Positive Controls

29. Do you have any genes designated as positive controls on your array? You typically have positive controls when there is DNA from a different genome than the one you are investigating on your array, and you added a known quantity of that DNA to your sample. Entering "true" as the object-value of the line given below means you have positive controls, and you want GeneSpring to normalize your experiment using the positive control values. This normalization method takes the average signal intensities of all of the positive controls and divides each gene's signal intensity by that number, for more information about this normalization option see "Normalizing Options" on page G-1. If you do not want to normalize your experiment using positive controls, either do not enter the "NormalizePosControl : " line, or type "false" as the object-value.

```
NormalizePosControl : either true or false
```

```
NormalizePosControl : true
```

The required layout file for positive controls

30. If you do not have positive controls or if you are not using them to normalize your data, skip this question and the associated *.experiment* file entry. If you are using positive controls you must have a *layout* file, and a file specifying what the positive controls are, this second file must have the gene names of the positive controls written in a list, one gene per line. See section "The Layout file" on page K-2 for more information about these files. Specify the complete file name of the layout file with the line below.

```
Layout : complete name of the layout file, the
file name can be anything, with or without spaces
```

```
Layout : AffyYeastLayout4.txt
```


Copyright 1998-2001 Silicon Genetics

There are two normalization options requiring you to have a layout file; both use the same line to tell GeneSpring where to find the file. You should only have one layout file, and you should only enter the line, "Layout : name of layout file", once. You may have already entered this file, please refer to "The required .layout file for Region Specifications" on page -9.

31. If you do not have positive controls or are not using them to normalize your data, skip this question and the associated *.experiment* file entry. Sometimes something will go wrong with the positive controls and you will get very low values for all of them, which you will not want to use for normalization purposes. Indicate the minimum average the positive controls must have such that dividing each genes' control strength by the average of the positive controls will not artificially inflate the noise of the genes.

```
NormalizeMinRange : indicate the minimum average
allowable for the positive controls

NormalizeMinRange : 10
```

The number indicated in the example (10) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.

Normalizations: Each Sample to Itself

32. Do you want to normalize your data by making the median of all of your measurements 1, for each sample in your experiment? (If you have not already preformed normalizations on your data you generally want to use this normalization option.) For more information about this normalization option, see "Normalizing Options" on page G-1.

```
NormalizeNoControl : either true or false

NormalizeNoControl : true
```

33. If you are not normalizing each sample to itself, skip this question and the associated *.experiment* file entry. Sometimes something will go wrong with the experiment and you will get very low values for everything. Indicate the cut-off value by telling GeneSpring not to raise all of the control strength values up to a median of 1 if their average is below this number:

```
NormalizeMinRange : Specify the cut-off value
telling GeneSpring not to raise all of the control
strength values up to a median of 1 if the average
control strength is below this number

NormalizeMinRange : 10
```

The number indicated in the example (10) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.

 Copyright 1998-2001 Silicon Genetics

Normalizations: Each Gene to Itself

34. Do you want to normalize each gene to itself, so the median of all of the measurements taken for the gene is one? See "Normalizing Options" on page G-1 for more information about this option. If you are not doing a two-color experiment you generally want to do this.

```
NormalizeEachGene : either true or false

NormalizeEachGene : true
```

35. Skip this question and the associated entry if you are not normalizing each gene to itself. Sometimes something will go wrong with the samples and all of the values for a particular gene are very low, in which case GeneSpring will artificially inflate the noise of the gene if you normalize those values up to a median of one. To specify where this cut-off is, type the line below in the *.experiment* file:

```
NormalizeMinMedian : the numerical cut-off value
below which you will not normalize a gene to
itself
NormalizeMinMedian : 0.01
```

The number indicated in the example (0.01) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.

Normalizations: Each Sample to a Specific Sample

36. Do you want to normalize each sample to one sample within the experiment? If so, enter the number of the sample, counting from zero as the object-value in the line below. Silicon Genetics does not recommend suggest using this normalization option, unless you have very specific reasons as described in "Normalizing Options" on page G-1.

```
NormalizeToExperiment : true or false

NormalizeToExperiment : 0
```


Copyright 1998-2001 Silicon Genetics

Colorbar Specifications

37. The intensity of the colorbar in GeneSpring indicates how reliable the data for each gene is. Indicate a raw control strength value to be considered very reliable (a high control strength) value, an average (a medium control strength) value, and an unreliable (a low control strength) value. Any gene with a control strength (control) above the value indicated as a high control strength will be colored using the brightest color appropriate, any gene with a control strength below the value given for unreliable data will be almost black in color. The medium signal value gives the value for the mid-point of the color bar, and genes with a medium control strength are colored halfway between the two color extremes. The default values are specified in the example. If you do not indicate a high, medium, and low values specifically, then the values GeneSpring will automatically use to determine the color bar are:

```
SignalHigh : a high number, this indicates high
confidence in the data
SignalMedium : a medium number, this indicates
average confidence in the data
SignalLow : a low number, this indicates low con-
fidence in the data

SignalHigh : 500
SignalMedium : 150
SignalLow : 50
```

These numbers are arbitrary. They are intended to be general benchmarks, not hard boundaries.

Graph Specifications

The values indicated here can be altered within GeneSpring, you are simply setting the default values here.

38. To allow you to inspect the genes' expression profiles closely, GeneSpring does not graph the entire y-axis (the expression level axis), but only the portion most genes profiles fall into. Indicate the range of expression levels GeneSpring should graph.

```
LowerBound : Indicate the lowest expression level
to graph on the y-axis
UpperBound : Indicate the highest expression level
to graph on the y-axis

LowerBound : 0
UpperBound : 5.0
```

A lower bound of 0 and an upper bound of 5 are the default settings of GeneSpring.

Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Appendix K Experiment File Formats

You can install a new experiment in one of several ways: by using the Experiment Installation Wizard (see "The Experiment Wizard" on page D-1) or by creating a *.experiment* file by hand (see "Installing from a Text File" on page J-1). Both experiment entry methods may involve a number of corollary files.

Only one file type is necessary for installing an experiment:

- Experimental data file(s), containing the genes' names and raw data for each sample in the experiment. Please refer to "Raw Data" on page -1.

Other helpful files might include:

- Layout file
- Region designation files
- A file listing the positive controls
- A file listing the negative controls
- GIF or JPEG pictures to be associated with this experiment, or with particular samples within the experiment
- GIF or JPEG pictures of the Microarray plates the experiment was done on

Raw Data

An experimental file consists of a list of gene names, a list of the raw data associated with them, and the condition or conditions during the test. In addition, an experiment may involve more than one sample, various normalization controls (such as positive and negative controls, and control channel values), pictures of the conditions during the experiment, and pictures of the array plates the experiments were done upon.

Copyright 1998-2001 Silicon Genetics

What format does this data need to be in?

Data may be in any of the following eight formats, depending on the type of data represented.

Experimental Data

You will need at least one file containing your experimental data. This file must have the gene names listed in one column, one name per line, with the experimental data reported in columns. If it were viewed in a spreadsheet it might look like this:

Gene Name	Control Strength in Experiment 1	Control Channel Strength	Background Signal	Background Signal for the Reference	Experiment Flag	Region
CLN1	510	110	10	10	P	A
MEP2	9	19	9	9	M	C

If created in a spreadsheet program, the file should be saved as a tab-delineated text file.

If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use commas and periods interchangeably.

GeneSpring can also read experimental data from databases via an ODBC link. Please refer to “Installing from a Database” on page E-1.

Pictures of the conditions during the experiment

At most there can be one picture associated with each condition. You do not need to have any pictures but they are good mnemonics, reminding you of what was happening in the experiment at the point you are viewing in GeneSpring. If you have only a few pictures, this can be very useful as GeneSpring will use the picture closest to the displayed condition. These pictures should be either GIF or JPEG files.

Pictures of the Microarray plates

At most there can be one array picture associated with each sample. They are helpful but not necessary. These pictures should be either GIF or JPEG files.

The Layout file

If you load experiments via the Experiment Wizard or the AutoLoader then you will probably never have to create your own layout file and thus you can skip this entire section. However, if you use the pasting option you may need to create the positive and/or negative control files associated with the layout file.

The layout file tells GeneSpring where to find other files associated with the experiment. If you load in experiments using a *.html* file, then you will need to create a layout file if each sample in

Copyright 1998-2001 Silicon Genetics

your experiment involved more than one array, and/or if the experiment used positive or negative controls. Frequently, the same layout file can be used for more than one experiment.

There are four possible lines in a layout file. Each line is either blank or a line of the form object-name space-colon-space object-value:

```
Object-name : object-value
```

An example of this is:

```
IncludePosControls : false
```

Here "IncludePosControls" is the object-name and "false" is the object-value. The object-value can be thought of as the answer to the question posed by the object-name. In the layout file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Usually when an experiment looks like it is not installed correctly it is because of a spelling or capitalization error. Using the copy (Ctrl+C) and paste (Ctrl+V) functions will help prevent this type of error.

This section is designed to help you create a layout file for a particular experiment, rather than explaining exactly what each possible answer means. There are two examples following each question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (**bold**-faced) example is an example of an actual answer to the question. A complete layout file for the fictitious "Yeast extraterrestrial studies" experiment is given at the end of this chapter.

The four possible lines in the layout file are:

1. Include this line if your experiment has positive controls. This line refers to a file listing the positive control. If you have positive controls you must have a separate file designating them. See "The Positive and Negative Control Files" on page -7 for information about this file.

   ```
   PosControlFilename : the complete file name of the
   file listing the gene names of the positive con-
   trols, one per line
   ```

   `PosControlFilename : PosControls.txt`

2. Include this line if your experiment has positive controls. This line tells GeneSpring if you want to display the positive control genes in the genome browser with the rest of the experiment, as if they were genes from the organism you are studying. Type "true" as the object-value for this line if you wish to view the positive controls in the genome browser, and enter "false" if you do not.

   ```
   IncludePosControls : true or false
   ```
   `IncludePosControls : false`

Copyright 1998-2001 Silicon Genetics

3. Include this line if your experiment has negative controls. This line refers to a file listing the negative control. If you have negative controls you must have a file designating them. See "The Positive and Negative Control Files" on page -7 for information about this file.

```
NegControlFilename : the complete file name of the
file listing the gene names of the negative con-
trols, one per line
```

```
NegControlFilename : NegControls.txt
```

4. Include this line if a sample in your experiment involved more than one array, or if there is some reason to normalize the sections of the array separately. If the genes from a sample could belong to more than one region, then the region must be noted somehow in the experimental data file (see "The Region Designation File(s)" on page -4). Use this line if the region is noted as either a unique entry in its own column or if it is a suffix appended to another column's entry. The object-value(s) in this line refer to separate files, each listing one possible region designator. See "The Region Designation File(s)" on page -4 for more information. Multiple region designation files should be separated with semicolons, but not spaces.

```
Regions : the complete file names of the files
listing the region designations, separated by
semicolons
```

```
Regions : YeastRA.txt;YeastRB.txt;Yeas-
tRC.txt;YeastRD.txt
```

The Region Designation File(s)

If there is more than one region to which the genes from a sample could belong, then the region must be noted somehow in the experimental data file. If the region is noted in the experimental data file as either a unique entry in its own column or as a suffix appended to another column's entry (as is common with Affymetrix chips) then you should create separate region designation files, one for each region. In this region designation file should be one line, reading:

```
RegionSuffix : character or string of characters
used either as a unique column entry or as a suf-
fix. This string designates a particular region.
```

```
RegionSuffix : A
```

All of the entries in the region column (designated in the *.html* file or in the "Regions Normalization" panel of the Experiment Wizard) having the same suffix as the object-value indicated after one of the "RegionSuffix : " entries are considered to be in the same region. For example, if there are four regions, A, B, C, and D there will be four region designation files, each with one of the lines:

```
RegionSuffix : A
RegionSuffix : B
RegionSuffix : C
RegionSuffix : D
```

 Copyright 1998-2001 Silicon Genetics

Given a region column in the experimental data file containing these entries:

```
Gene1A
Gene2B
Gene3C
Gene4D
Gene5A
Gene6B
Gene7C
Gene8D
Gene9A
. . .
```

In this example, genes 1, 5, and 9 are all marked as in region A and could be normalized as a discrete group.

An Example:

You have experiment 1 with subchips A, B, C, Da, Dd (2 repeats for subchip D) to be compared to experiment 2 with subchips A, B, Ca, Cb, D (2 repeats for subchip C). You can load it as four samples.

Exp 1:	A	B	C	Da
Exp 2:				Db
Exp 3:	A	B	Ca	D
Exp 4:			Cb	

Table A-1 Correct entry of repeated sub-experiments

Give experiments 1 and 2 the same parameters. Give experiments 3 and 4 the same parameters.

Entering region specifications when they are not specified in their own column or as suffixes within another column

Copyright 1998-2001 Silicon Genetics

Occasionally a region may not be designated by a unique column entry or as a suffix appended to a column entry. In this case you cannot use the Experiment Wizard to automatically read in your region designations. You will need to create a layout file for your experiment and separate region designation files. A region designation file is used to describe a region, and specifies the following information:

- How to distinguish this region from other regions.
- How to map gene names in this region to the gene names given in the list of genes defining the genome.

There are several ways regions can be distinguished. The four ways listed below are typically used separately, but can occasionally be used in combination, with each other or with the standard way to designate a region.

1. The regions are defined implicitly by the order the genes names as reported in the experimental data file. The names of the genes can be sorted in alphabetical order and used to determine whether a gene is in this region. One can specify inclusive beginning and ending genes, and any genes between them (alphabetically) will be considered part of this region. See the next option for the meaning of "UsesCommas".

```
EndRegion : the last gene name in the region
StartRegion : the first gene name in the region
UsesCommas : false

EndRegion : s191
StartRegion : s001
UsesCommas : false
```

2. The regions are defined implicitly by the ordered names of the genes, in a rectangular coordinate system. This is similar to the previous option, except the "names" of the genes are actually coordinates, separated by commas. In this case, a gene is only in the given region if it is between the starting and ending gene names for each dimension separated by commas. For instance:

```
StartRegion : 001,100
EndRegion : 099,199
UsesCommas : true
```

3. The regions are defined explicitly by a list of gene names, and optionally a change of names. In this case, you must define a map for the region. A map can be just a list of genes, or it can be a list of names (as used in the experiment files) and the corresponding gene names (as used in gene list defining the genome). In this case, you must specify a text file describing the map (see "How to describe a map" on page -7).

```
Map : mapA.txt
```

 Copyright 1998-2001 Silicon Genetics

4. The regions are defined by file name extension. The experimental data for each region is in a separate file. The file names for each sample specified in the Experiment Wizard or in the *.html* file are base names, and each region adds an extension to this file name. To prevent name conflicts, this option is frequently used with the map option.

```
FileNameExtension : .chipA
```

How to describe a map

Maps are used when you want to change gene names from the raw names (e.g. chip coordinates) into more standard gene names. They can also be used to specify a list of genes defining a region. A map file is a text file containing just two lines:

```
FileName : GeneList.txt
ChangeNames : true
```

The "FileName" entry specifies the name of a text file containing one line per gene. If "ChangeNames" is true, then the text file should consist of two columns (separated by a tab). The first column should be the gene names as they appear in the experiment data file; the second column should be the gene names as they appear in the list of genes defining the genome. If "ChangeNames" is false, then the text file should only have one column. In this case, the map is used only to specify what is present in a region.

The Positive and Negative Control Files

A positive control file and a negative control file are formatted in exactly the same way; their contents are different. Each file lists the control genes' names, one name per line:

```
Control Gene Name 1
Control Gene Name 2
Control Gene Name 3
Control Gene Name 4
Control Gene Name 5
Control Gene Name 6
. . .
```

This list of gene names is all either file should contain. There should not be any headlines or anything else in the file, only the gene names.

Briefly, you have negative controls in your experiment when there is DNA from a different genome than the one you are investigating on the array. You are using positive controls when there is DNA from a different genome than the one you are investigating on your array, and you add a known quantity of that different DNA to your sample. For a description of the possible normalizations to be done with these controls see "Normalizing Options" on page G-1.

The names of the positive and negative controls do not need to be listed in your Master Table of Genes. If they are listed, those genes will be colored gray (not measured) in the genome browser because they are used in normalization not measurement.

 Copyright 1998-2001 Silicon Genetics

Where do I put my data?

There are eight possible raw data files listed below; only the first one is necessary for loading an experiment.

You must have:

- Experimental data file(s), containing the genes' raw data for each sample in the experiment. Please refer to "Raw Data" on page -1.

You might have:

- A Layout file
- Region designation file(s)
- A map file
- A file listing the positive controls
- A file listing the negative controls
- GIF or JPEG pictures of the conditions during the experiment
- GIF or JPEG pictures of the Microarray plates the experiment was done on

All of the raw data files should all be placed within the "Experiment" sub-folder of the organism they pertain to. The default pathway for this directory is:

```
C:/Silicon Genetics/GeneSpring/Data/Genome Name/
Experiments
```

If the defaults were changed, your version of GeneSpring may be stored elsewhere, but the end of the pathway should be identical on your computer.

Copyright 1998-2001 Silicon Genetics

Appendix L Equations for Correlations and other Similarity Measures

Many of the advanced analysis technics are based upon measures of gene similarity. Similarity or "nearness" between genes is usually based on the correlation between the expression profiles of the two genes. GeneSpring offers nine choices of similarity measures. Each is selectable from a drop-down list appearing the Clustering and Filtering windows. Please refer to Chapter 5, Clustering and Characterizing Data in GeneSpring and "Filter Genes Analysis Tools" on page 4-1 repectivily.

Each measure takes two expression patterns and produces a number representing how similar the two genes are. Most of the measures of similarity are correlation measures, and their value will vary from -1 (exactly opposite) to 1 (the same). For a measure of distance, the result will vary from 0 (the same) to infinity (different). For confidences, the result will vary from 0 (no confidence) to 1 (perfect confidence). Both distance and confidence are actually measures of dissimilarity (small means close and large means far away). These are each transformed to measures of similarity by GeneSpring in ways detailed below.

If one expression value for a particular experiment for either gene is missing, that experiment will be not considered in the calculation.

The notation used to describe the formulas:

- *Result* : the result of the calculation for genes **A** and **B**.
- n : the number of samples being correlated over.
- **a** : the vector ($a_1, a_2, a_3 \ldots a_n$) of expression values for gene **A**.
- **b** : the vector ($b_1, b_2, b_3 \ldots b_n$) of expression values for gene **B**.

Normal mathematical notation for vectors will be used. In particular:

- $\mathbf{a.b} = a_1b_1+a_2b_2+\ldots+a_nb_n$
- $|\mathbf{a}|$ = square root($\mathbf{a.a}$)

Copyright 1998-2001 Silicon Genetics

Common Correlations

Standard Correlation

Standard correlation measures the angular separation of expression vectors for Genes **A** and **B** around zero. As almost all normalized values for genes are positive, you find mostly positive correlations between genes when you use the Standard correlation. This metric is designed to answers the question "do the peaks match up?" or to put it another way, "are the two genes expressed in the same samples?" Since these questions are the most frequent questions a biologist is trying to get answered, GeneSpring calls it "Standard correlation". It is important to note, what mathematicians and statisticians refer to as "correlation" usually refers to the Pearson correlation. The "Standard correlation" would be called "Pearson correlation around zero" by mathematicians and statisticians.

This is how to compute a Standard correlation:

Standard correlation = **a.b**/*(*|**a**||**b**|*)*

Pearson Correlation

The Pearson correlation is very similar to the Standard correlation, except it measures the angle of expression vectors for genes **A** and **B** around the mean of the expression vectors (for example, the mean of the expression values constituting the profiles for Gene **A** and Gene **B**). Generally the mean of the expression vectors will be positive since expression values are based on concentrations of mRNA. Using the Pearson correlation you get more negative correlations then you get from the Standard correlation (for example, you find more genes that behave opposite to each other, because of where you put the baseline—at zero almost all gene values are above it, at 1 there are a fair amount that read below the baseline). It is worth noting that, for data normalized to an overall level of 1 (as with all normalizations that GeneSpring performs) the Pearson correlation gives you almost the same correlations as the Standard correlation when they are both performed on the logarithms of the genes' expression values.

This is how to compute a Pearson Correlation:

Calculate the mean of all elements in vector **a**. Then subtract that value from each element in **a**. Call the resulting vector **A**. Do the same for **b** to make a vector **B**.

Pearson Correlation = **A.B**/*(*|**A**||**B**|*)*

Copyright 1998-2001 Silicon Genetics

Spearman Correlation

The Spearman correlation is a nonparametric correlation similar to the Pearson correlation except it replaces the data for Gene **A** and **B** with the ranks of the data (i.e. the lowest measurement for a gene becomes 1, the second lowest 2, and so forth). Spearman correlation calculates the correlation of the ranks for Genes **A** and **B**'s expression data around the mean of the ranks, using the same formula as Pearson correlation. In the Spearman correlation only the order of the data is important, not the level, therefore extreme variations in expression values have less control over the correlation. If there are ties in the data, then all of the tied values are assigned the average of the ranks, e.g. if the 5th, 6th and 7th lowest values are tied, all three datapoints are assigned a rank of 6.

This is how to compute a Spearman correlation:

Order all the elements of vector **a**. Use this order to assign a rank to each element of **a**. Make a new vector **a'** where the i^{th} element in **a'** is the rank of a_i in **a**. Now make a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, make a vector **B** from **b**.

Spearman correlation = $\mathbf{A}.\mathbf{B}/(|\mathbf{A}||\mathbf{B}|)$

Spearman Confidence

Spearman confidence is a measure of similarity, not a correlation. Spearman confidence is one minus the p-value for the statistical test when the Spearman correlation is zero versus the alternative when it is larger than zero. There is a high Spearman confidence value if there is a high Spearman correlation and a low p-value, meaning there is a low probability to find a correlation this high. This measure is very similar to looking for large Spearman correlation values, but it takes account of the number of sub-experiments in your experiment set.

This is how to compute a Spearman confidence:

If *r* is the value of the Spearman correlation as described in "Spearman Correlation" on page -3, then:

Spearman confidence =1-(probability you would get a value of *r* or higher by chance.)

Two-sided Spearman Confidence

Two-sided Spearman confidence is again a measure of similarity but not a correlation. It is very similar to the Spearman confidence discussed in "Spearman Confidence" on page -3, except it is based on the two-sided test of whether the Spearman correlation is either significantly greater than zero or significantly lower than zero. There is a high Two-sided Spearman confidence value if the absolute value of the Spearman correlation is large and has a small p-value, meaning there is a low probability to find a correlation with absolute value this large.

This "similarity" measure is really good for answering the question "What genes behave similarly to a specific gene, and at the same time, what genes behave opposite to a specific gene?". It should probably not be used for the advanced clustering algorithms (such as k-means and hierarchical clustering) because the genes with high two-sided confidence values are really a mixture of similar and dissimilar genes.

Copyright 1998-2001 Silicon Genetics

This is how to compute a Two-sided Spearman confidence:

If *r* is the value of the Spearman correlation as described above, then:

Two-sided Spearman confidence =1-(probability you would get a Spearman correlation of |*r*| or higher, or -|r| or lower, by chance.)

Distance

Distance is not a correlation at all, but a measurement of dissimilarity. Distance is based on the measurement of Euclidian distance between the expression profile for gene **A** (defined by its expression values for each point in N-dimensional space, where **N** is the number of experimental points (conditions) with data in your experiment) and the expression profile for gene **B**. This is more formally known as the Euclidian metric. To standardize this difference GeneSpring divides by the square root of the number of conditions.

This is how to compute a Euclidian Distance:

Distance = |**a-b**| /square root of N
Since distance is a measure of dissimilarity, the distance (d) is converted when needed to a similarity measure 1/(1+d).

Special Case Correlations

The next three metrics should only be used to look at special cases. They are all modified versions of the Standard correlation. Using these three metrics only makes sense when your data is in a sequence, such as “before” and “after”, a time series, or a drug series. The sequence does not have to be continuous, but it must have an order. If your experiment is set up with an experimental point taken at each of “before”, “after”, and “control” then the following correlations will not make sense applied to your data.

Smooth Correlation

This is how to compute a Smooth correlation:

Make a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph screen. Do the same to make a vector **B** from **b**.

Smooth correlation = **A.B**/*(*|**A**||**B**|*)*

Copyright 1998-2001 Silicon Genetics

Change Correlation

The Change correlation looks for the opposite of what the Smooth correlation looks for. The change correlation only looks at the change in expression level of adjacent points. However, it is also very similar to the Standard correlation, in that it measures the angular separation of expression vectors for genes A and B around zero (i.e. in comparison to zero), except instead of using the expression values in each experimental point to create the expression vector for gene A, it is based on an arc tangent transformation of the ratio between adjacent pairs of experimental points and uses these to create the expression vector. This correlation looks for when gene A and gene B are changing at the same time. Using the arc tangent makes a measure of change that is less sensitive to outliers than using the ratio directly.

This is how to compute a Change correlation:

Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i)-\pi/4$.Do the same to make a vector **B** from **b**.

Change correlation = $\mathbf{A}.\mathbf{B}/(|\mathbf{A}||\mathbf{B}|)$

Upregulated Correlation

The Upregulated correlation is very similar to the Change correlation, except that it only considers positive changes. All negative values for the arc tangent transform of the ratio are set to zero. This emphasizes only periods when new RNA is being synthesized.

This is how to compute an Upregulated correlation:

Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\max(\text{atan}(a_{i+1}/a_i)-\pi/4,0)$. Do the same to make a vector **B** from **b**.

Upregulated correlation = $\mathbf{A}.\mathbf{B}/(|\mathbf{A}||\mathbf{B}|)$

Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Appendix M Creating an Array in GeneSpring

In order to create an array layout file in GeneSpring, you need at least one file to tell GeneSpring general information about the array (size, shape, features, format, name, etc.). This file should end in the extension *.layout*. You usually need another file describing exactly which gene goes where.

The format of the *.layout* file is a series of lines (order does not matter). Each line consists of a property, a colon, and a value. For example, `property : value`. Blank lines and lines starting with a number sign (#) are ignored by GeneSpring. The following properties are allowed in the file. As always, GeneSpring is case-sensitive, so please use the capitalizations as presented here:

- **Name:** The name of this layout, to appear in the navigator window of GeneSpring.
- **Icon:** (optional) The path of a 16 by 16 *.gif* file to appear next to the layout in the navigator window.
- **VerticalSubArrays:** (optional, default 1) The number of rows of sub-arrays.
- **HorizontalSubArrays:** (optional, default 1) The number of columns of sub-arrays.
- **HorizontalPerSubArray:** The number of columns of dots in a sub-array.
- **VerticalPerSubArray:** The number of rows of dots in a sub-array.
- **VerticalDuplication:** (optional, rarely used) When dots are duplicated vertically, the number of copies.
- **HorizontalDuplication:** (optional, rarely used) When dots are duplicated horizontally, the number of copies.
- **CommonArrayType:** The format of the array.
    - **Q-X-Y**—The data file contains two columns. The first is a list of genes, the second is a set of three numbers separated by commas or hyphens. The first is the "sub-array" number, the second is the X-coordinate, and the third is the Y-coordinate. All numbers start counting from 1. The subarrays are counted left to right, top to bottom. The second column can optionally be enclosed in quotation marks.
    - **Q-R-C**—Same as "Q-X-Y", except the X and Y coordinates are swapped.
    - **CLONTECH LNL**—There is no datafile. All genes have systematic names of the form "B4c" indicating where they are in the array. The first (capital) letter indicates which sub-array; the number indicated which column, and the lower case letter indicates which row.
- **CLONTECH LNNL:** Same as LNL, except there are two digits instead of one.
- **DataFileName:** The name of a datafile linking locations with gene names in format given by the CommonArrayType choice. In the second example below there are several lines of a DataFile file.

Copyright 1998-2001 Silicon Genetics

Once you are done creating the *.layout* file you should save it in the ArrayLayouts folder of the genome folder for which the layout pertains. For example, if you have not changed the defaults set-up of GeneSpring the path to the layout folder in the yeast genome would be C:\Program Files\SiliconGenetics\GeneSpring\data\yeast\ArrayLayouts.

Examples of *.layout* files for Arrays

Here is an example for Pat Brown's yeast layout. The following is from a file *Pat.layout*:

```
Name : Pat Brown's Yeast Layout
# Icon : XXX.gif
VerticalSubArrays : 2
HorizontalSubArrays : 2
HorizontalPerSubArray : 40
VerticalPerSubArray : 40
VerticalDuplication : 1
HorizontalDuplication : 1
CommonArrayType : Q-X-Y
DataFileName : PatLocationList.txt
```

Following are the first few lines of the file *PatLocationList.txt*:

```
YHR007C "1,13,1"
YBR218C "2,13,1"
YAL051W "1,14,1"
YAL053W "2,14,1"
YAL054C "1,15,1"
YAL055W "2,15,1"
YAL056W "1,16,1"
```

Here is an example for a CLONTECH Array, from a file *Clontech.layout*:

```
Name : Clontech 588
# Icon : XXX.gif
VerticalSubArrays : 2
HorizontalSubArrays : 3
HorizontalPerSubArray : 14
VerticalPerSubArray : 14
VerticalDuplication : 1
HorizontalDuplication : 2
CommonArrayType : Clontech
```

Making an array is a complicated process, please contact Silicon Genetics' Technical Services Department at 650-367-9600 or support@sigenetics.com for more information on this topic.

Copyright 1998-2001 Silicon Genetics

Appendix N Technical Details on the Statistical Group Comparison

Statistical Group Comparison is a filter tool that statistically compares mean expression levels between two or more groups of samples. The object is to find the set of genes for which the specified comparison shows statistically significant differences in the mean normalized expression levels as interpreted according to your current interpretation mode (logarithm, ratio or fold change) across all the groups[1]. This comparison is performed for each gene, and the genes with the most significant differential expression (smallest p-value) are returned. The comparisons can be done with parametric or non-parametric methods. The parametric comparison for two groups is known as Student's two-sample t-test. For multiple groups, this is known as one-way analysis of variance (ANOVA). You can specify whether to assume within-group variances are equal across all groups. Calculations without the assumption of equality of variances are done using Welch's approximate t-test and ANOVA. Non-parametric comparisons are also available, corresponding to the Wilcoxon two-sample text (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups.

For Each Gene

For each gene separately, GeneSpring will do the following:

Let i index over the G groups formed by distinct levels of the comparison parameter. Let X_{ik} be the expression values, with k running over the replicates for each situation, interpreted according to the current interpretation (ratio, log of ratio, fold change). Let

N_i = the number of non-missing data values for each group

$\overline{X_i} = 1/N_i \sum_{k=1}^{N_i} X_{ik}$ be the group means, and

$SS_i = \sum_{k=1}^{N_i} (X_{ik} - \overline{X_i})^2$ be the within-group sum of squares.

In all calculations here, missing (NaN) values are left out of the sums, not propagated.

If any of the N_i are zero, drop that parameter level from the analysis, and readjust G accordingly. If G is not at least 2, exit (p-value=1).

1. Filtering genes based on a one-sample t-test of the mean expression level across repeats or replicates versus a reference value can be done by selecting "t-test p-value" as the filter criteria in Expression Percentage Restriction.

Copyright 1998-2001 Silicon Genetics

Parametric Test, Variances Assumed Equal

For parametric test, with variances assumed equal, compute:

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{G} N_i \overline{X_i}}{\sum_{i=1}^{G} N_i} \text{ the overall mean,}$$

$$BSS = \sum_{i=1}^{G} N_i (\overline{X_i} - \overline{\overline{X}})^2 \text{ the between-groups sum of squares,}$$

$d_1 = G - 1$ the numerator degrees of freedom,

$BMS = BSS / d_1$ the between-groups mean square,

$WSS = \sum_{i=1}^{G} SS_i$ the pooled within-group sum of squares,

$d_2 = \sum_{i=1}^{G} (N_i - 1)$ the denominator degrees of freedom,

if d_2 is not greater than zero, then exit (p-value=1).

$WMS = WSS / d_2$ the within-groups mean square, and

$F = BMS / WMS$ the F-ratio statistic

if $WMS = 0$ then make F is treated as arbitrarily large (p-value = 0).

The p-value is calculated by looking up F in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom

Parametric Test, Variances Not Assumed Equal

For the parametric test without assuming variances equal:

First check that each group has N_i greater than or equal to 2 and SS_i greater than 0, if not, remove it from consideration and recompute G again. If G is not at least 2, exit (p-value=1)[1].

1. This reflects the more stringent requirements of not assuming the variances equal – if the variance estimate is pooled, replicates are only needed for at least one group, if variances are separately estimated then replicates are needed for each group.

Copyright 1998-2001 Silicon Genetics

Then compute:

$$w_i = N_i\left(\frac{N_i - 1}{SS_i}\right)^2 \text{ the group weights}$$

$$W = \sum_{i=1}^{G} w_i \text{ the sum of weights}$$

$$\tilde{X} = \frac{\sum_{i=1}^{G} w_i \overline{X_i}}{W} \text{ the weighted mean}$$

$$BSS = \sum w_i\left(\overline{X_i} - \tilde{X}\right)^2 \text{ the between-groups sum of squares}$$

$$d_1 = G - 1 \text{ the numerator degrees of freedom}$$

$$BMS = BSS/d_1 \text{ the between-groups mean square}$$

$$Z = \frac{1}{G^2 - 1}\sum_{i=1}^{G}(1 - \frac{w_i}{W})^2 \Big/ (N_i - 1)$$

$$d_2 = \frac{1}{3Z} \text{ the denominator degrees of freedom}$$

if d_2 is not greater than zero, then exit (p-value=1).

$$WMS = 1 + 2(G - 2)Z \text{ the within-group mean square}$$

$$W = BMS/WMS \text{ the test statistic}$$

The (approximate) p-value is calculated by looking up W in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom. Note that d_2 will not, in general, be an integer.

Nonparametric Analysis

For the nonparametric analysis:

Replace each X_{ik} by R_{ik}, their rank out of all of the $\{X_{ik}\}$ for the gene. Perform the same analysis as for parametric test with variances equal. P-values are approximate but asymptotically accurate.

Copyright 1998-2001 Silicon Genetics

References

Brown, M.B., and Forsythe, A.B. (1974) The small sample behavior of some statistics which test the equality of several means. *Technometrics* 16, 169-132.

Conover, W.J. (1980) Practical Nonparametric Statistics, 2nd Ed. New York, John Wiley & Sons, Inc.

Scheffe, H. (1959) The Analysis of Variance, New York: John Wiley & Sons, Inc.

Copyright 1998-2001 Silicon Genetics

Appendix O Technical Details for the Predictor

Gene Selection

In order to select genes for use in the predictor, all genes are examined individually and ranked on their power to discriminate each class from all others, using the information on that gene alone. For each gene, and each class, all possible cutoff points on gene expression level for that gene are considered to predict class membership either above or below that cutoff. Genes are scored on the basis of the best prediction point for that class. The score function is the negative natural logarithm of the p-value for a hypergeometric test (Fisher's exact test) of predicted versus actual class membership for this class versus all others.

A combined list containing the most discriminating genes for each class is produced as the predictor list. Each class is examined in turn, and the gene with the highest score for that class is added to the list, if it is not already on the list. Then genes with the next highest scores for each class are added. This is continued in rotation among the classes until the specified number of predictor genes is obtained. If you save the list of predictor genes as a Gene List, the best prediction score of the gene among the classes for which it would have been added to the list is saved as the attached number on the list.

Classifying the Test Samples

Based on the selected genes, classifications are then predicted for the independent test data, using the k-nearest-neighbors rule. A sample in the independent set is classified by finding the (user specified) k nearest neighbors of the sample among the training set samples, based on Euclidean distance between the normalized expression ratio profiles of the samples. The class memberships of the neighbors are examined, and the new sample is assigned to the class showing the largest relative proportion among the neighbors after adjusting for the proportion of each class in the training set.

Decision Threshold

P-values are computed for testing the likelihood of seeing at least the observed number of neighborhood members from each class based on the proportion in the whole training set. The class with the smallest p-value is given as the predicted class. The column labeled "P-value" is the ratio of the p-value for the best class to that of the second-best class. The predictor will make a prediction if this ratio is less than the "P-value Cutoff" specified on the initial panel, and will not make a prediction if the ratio is above this cutoff. Setting the p-value cutoff to 1 will force the algorithm to always make a prediction but may result in more actual prediction errors.

Copyright 1998-2001 Silicon Genetics

References for the Predictor

Cover, T.M. and Hart, P.E. (1967) “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, IT-13, 21-27.

Duda, R. O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis, Wiley, New York.

Golub, T.R. et. al. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring” Science, v286, pp 531-537 (1999)

Copyright 1998-2001 Silicon Genetics

Appendix P Common Commands

There are a number of common commands available in nearly all of the GeneSpring screens. Not every command listed here will be available in every screen, nor is every command available listed. Commands specific to particular displays will be described in greater detail in those chapters.

Commands Accessible by Cursor or Keyboard

- **Select:** You can select a gene by clicking it. You can select more than one gene by clicking subsequent genes while holding the shift button down. You can select all the genes in an area by left-clicking in one corner of a rectangle and dragging to the opposite corner, while holding down the Shift key. If you know the systematic or common name of your gene, you can select it by selecting **`Edit > Find Gene`**.
- **Gene Inspector:** Double-click any gene in the browser to bring up the Gene Inspector. Or, if a gene is already selected, you can use **`Edit > View details on selected gene`** command or type **Ctrl+I**. This command brings up a window with more detailed information about a particular gene. For more information, see the "Gene Inspector" on page 3-37. Close the Gene Inspector by clicking the **`Cancel`** button.
- **Zoom In:** This command allows you to have a closer look at a particular section or point within the browser. Zooming is accomplished by clicking in the upper left corner of the region you wish to enlarge, and dragging the cursor to the lower right corner. Repeat until the desired magnification is reached. Systematic and then common gene names (if they exist) are listed beneath the gene as soon as there is adequate space under their associated rectangle. Sequence information is not visible in the Gene Inspector.
- **Arrow Keys:** When the genome browser is magnified by Zooming, the arrow keys on the keyboard allow you to shift the particular section being displayed in the direction of the arrow pressed.
- **Page up/Page Down:** Like the arrow keys, except over a larger scale, the Page Up/Page Down keys on a typical keyboard allow you to vertically pan through the genome browser.

Copyright 1998-2001 Silicon Genetics

Common Commands in the Drop-Down menus

The File Menu

- **Print:** You have several options on how to print from GeneSpring or save graphics as a file.
- **New Genome or Array:** This command will allow you to select from a submenu of available genomes. Selecting will bring up a new main GeneSpring window with your chosen genome displayed.
- **New Pathway:** This command will bring up the New Pathway Wizard. Please see "Pathways" on page 4-23 for more details.
- **Save Bookmark:** A Bookmark will save your analysis at its current point so you can come back to it later. Save your bookmark by selecting **`File > Save Bookmark`**. You will need to input a name for the bookmark. To open your saved Bookmark, go to the Bookmark folder and select a bookmark to view.

The File drop-down menu also gives you several options for loading genomes and experiments into GeneSpring, please refer to the *GeneSpring Loading Data Manual*.

The Edit Menu

- **Copy:** The copy menu allows you to copy gene lists, experiments or fully annotated gene lists to the clipboard, if the experiments are properly set up. Please refer to "Copying and Pasting Experiments" on page F-1 for more details.
- **Paste:** The paste menu allows you to insert an entire experiment from the clipboard, if the experiment is properly set up. Please refer to "Copying and Pasting Experiments" on page F-1 for more details.
- **Find Gene:** A particular gene can be found directly by using **`Edit > Find Gene`**; type either the systematic or the common name in the Find Gene box, then click **`OK`** or depress the **Enter** key. The genome browser will be zoomed around your selected gene. You can also type in a keyword such as "immun" and GeneSpring will present you with a list of genes and allow you to select one by clicking on the name, or save the list as a gene list. You can also bring up the Find Gene window by typing **Ctrl+F**.
- **Undo: `Edit > Undo`** will undo your last action. The Undo command has some memory, so you may be able to undo several actions. You can also Undo by typing **Ctrl+Z**.
- **Preferences:** This window will allow you to change many of the default settings in GeneSpring, including the colors used to display the genes. For more information, please refer to "Preferences Window" on page B-1.

Copyright 1998-2001 Silicon Genetics

The View Menu

In the View menu are all the display options you may choose for your data.

- **Unsplit Window:** The Split Window command allows you to view multiple graphs simultaneously in the genome browser. To split the window, right-click over a Gene Lists folder or a classification in the navigator and select **Split window** from the pop-up menu. Unsplit Window allows you to undo that feature and return to a normal screen.
- **Visible:** Under this command is a submenu presenting you with what you may show in your current view. If you are trying to maximize the screen, you can turn all the options off.

The Experiments Menu

- **Merge/Split Experiments:** This command allows you to merge data from several experiments into one or to split data from one experiment into several. Please refer to "Merging, Splitting and Duplicating Experiments" on page 2-6 for more information. You can also use this feature to copy an experiment.
- **Change Experiment Parameters:** This command allows you to add new, change or delete various parameters from your experiment. Please refer to "Normalizing Options" on page G-1 for more information.
- **Experiment Normalizations:** This command allows you to change the normalization technique used on your experiment. For an overview of the possible normalizations, please refer to "Normalizing Options" on page G-1.
- **Change Experiment Interpretation:** With this command you can change various aspects of the displayed experiment; for more details please see "Changing the Experiment Interpretation" on page 2-17.

The Colorbar Menu

You can change any of the default colors used in the genome browser. For more information, please refer to "Preferences Window" on page B-1. You can also right-click over the colorbar to change the range of brightness (trust) of the colors.

- **Color by Expression (Current Experiment):** Selecting the first command in the list will return you to the default coloring for your current experiment. Please refer to "Color by Expression" on page 3-31 for more details on this topic.
- **Color by Significance:** Please refer to "Color by Significance" on page 3-33 for more details on this topic.
- **Venn Diagram:** This command allows you to assign various gene lists to colored circles within a Venn Diagram. The submenu contains three options: left, right and bottom. Please refer to "Color by Significance" on page 3-33 for more details on this topic.
- **Color by Parameter:** This option allows you to color your genes by any parameter set as color code in the current interpretation. Please refer to "Color by Parameter" on page 3-33 for more details on this topic.

Copyright 1998-2001 Silicon Genetics

- **Color by Classification:** This command allows you to color all the genes by a classification. Please refer to "Color by Classification" on page 3-34 for more details on this topic.

The Tools Menu

- **Filter Genes:** This command allows you to make specific lists of genes according to their expression levels or other data. Please refer to the Chapter 4, Analyzing Data in GeneSpring for more details.
- **Clustering:** The Clustering command opens a new Cluster window. In the middle of the Cluster window is the Clustering Method drop-down menu in which you can choose one of the following clustering methods:
  - K-means: For more information, see "k-Means Clustering" on page 5-9.
  - Trees: This window allows you to create new gene trees or experiment trees. For more information, see "Trees" on page 5-1.
  - Self-Organizing Map: For information on Self-Organizing Maps (SOM), please refer to "Self-Organizing Maps" on page 5-12 or contact Silicon Genetics' technical service department at support@sigenetics.com or call 650-367-9600.
- **Show Drawable Gene:** This command will bring up the straight line of a manipulatable pseudo (drawn) gene. Please refer to "Creating Drawn Genes" on page 4-22 for more information.
- **Find Interesting Genes:** This function finds genes with the greatest trust values who go through the largest expression changes during the experiment. Please refer to "Find Interesting Genes" on page 4-21 for more information.
- **Find Potential Regulatory Sequences:** This command initiates the Find Potential Regulatory Sequence window, which allows you to specify certain parameters for an oligomer search in the nucleotide sequence preceding the genes in the list being displayed in the genome browser, and to perform the search. For more information about this window see "Regulatory Sequences" on page 4-26. If the nucleotide sequence has not been loaded a window will temporarily appear saying, "Please wait while the nucleic acid sequence is being loaded".
- **Principal Components Analysis:** For information on Principal Component Analysis (PAC), please refer to "Principal Components Analysis" on page 5-5 or contact Silicon Genetics' technical service department at support@sigenetics.com or call 650-367-9600.
- **GeneSpider:** This command will activate the GeneSpider. You can choose one of the available databases to update your information. The GeneSpider will do an automatic web search to see if anything new has been added to the public databases from which your information came.

Copyright 1998-2001 Silicon Genetics

Common Commands in the Genome Browser

Right-clicking in the genome browser will bring up a list of commands that can be performed from that window. Some of these commands are also available when right-clicking in the main screen of the Gene Inspector.

Mac Users should use Control-Click to activate pop-up menus.

- **Zoom Out:** Clicking the **`Zoom Out`** button or menu option (under View) will zoom out by a factor of two, as will **Ctrl+[**. You can also use **`Edit > Undo`** to go back to the previous level of magnification.
- **Zoom Fully Out:** This command returns the screen to its original magnification state (a magnification value of 1). Select **`View > Zoom Out`**. **Zoom Fully Out** is also in the menu resulting from a right-click while the cursor is in the genome browser. The **Home** key will also zoom the genome browser fully out.
- **Make List from Selected Genes:** This command allows you to make a new list from the genes highlighted in the genome browser. To use this command, right-click in the browser display window and a menu will appear. Go to the **Make List from Selected Genes** command and click it. A New Gene List window will appear. For more information about this window, see "New Gene List window" on page 4-11. If there are no genes selected, this command is disabled.

The Options Submenu

The **Options** submenu presented at the bottom of the right-click pop-up menu in the genome browser. It contains a number of possible options. Not all of these will be present, as many are dependent on the type of view selected. Most are simple toggle switches; simply select the same command again to turn it off.

Mac Users should use Control-Click to activate pop-up menus.

- **Change Vertical Axis Range:** You can use this command to change the upper and lower bonds of the vertical axis range. By using this command you can widen or compress the amount of information seen in the genome browser. Select **Change Vertical Axis Range** and the Parameter Bounds box will appear. Type in the new values and click **`OK`**. For more details, please refer to "To view a Scatter Plot" on page 3-16.
- **Load Sequence:** If you see this command, it is time to update your version of GeneSpring, as versions 4.0 and later load the sequence information automatically. Please refer to "Update GeneSpring" on page A-2 for details. If you have an older version, you can explicitly load sequences by right-clicking while the cursor is in the genome browser. A menu will appear. Go to the **Options** menu, and select the **Load Sequence** option. A window saying, "Please wait while nucleic acid sequence is loaded" will appear. After the loading is complete it is possible to zoom in and see the nucleic acid sequence of a particular gene. Loading the sequence also allows you to take advantage of GeneSpring's sequence-based features such as Find Regulatory Sequences.

Copyright 1998-2001 Silicon Genetics

- **Show ORF direction/Ignore ORF direction:** A gene is represented visually by a colored line or, upon higher magnification, a colored rectangle. The rectangle's position relative to the chromosome line determines the direction of the ORF. A gene below the chromosome line has a reading direction opposite to the direction chosen by the sequencers, and the sequence is read backwards. You can choose to display this distinction between which direction a gene is read (Show ORF direction) or to have no distinction between genes (Ignore ORF direction). Select the **Ignore ORF direction** command or the **Show ORF direction** command.
- **Show Complementary Bases/just Show One Strand Of Bases:** Show Complementary Bases allows both the Watson strand (5') and the Crick strand (3') to be shown while viewing the nucleic acid sequence in the physical position display, and conversely, Just Show One Strand Of Bases shuts this feature off and only displays the Watson strand of the sequence. Select the **Just Show One Strand Of Bases** command or the **Show Complementary Bases** command.
- **Show Horizontal Label/Hide Horizontal Label:** The horizontal axis is the experiment parameter. This command allows the label associated with the horizontal axis to be seen (or hidden.) The horizontal label is displayed in the bottom right corner of the Physical Position view. To hide this label, right-click while the cursor is in the genome browser. A menu will appear, go to the Options submenu, and select the **Hide Horizontal Label** option. To show this label, go to the same menu and select the **Show Horizontal Label**.
- **Show Vertical Label/Hide Vertical Label:** This feature allows the vertical label, which runs along the left side of the graph, to be seen or hidden. Normally in the Graph view, the vertical label is *Expression*. To hide this label, right-click while the cursor is in the genome browser. A menu will appear; go to the Options submenu, and click the **Hide Vertical Label** option. To show the vertical label, go to the same menu and click **Show Vertical Label**.
- **Label vertical axis on side/ Label vertical axis at top:** This feature is only applicable if the vertical axis label is visible. The label may appear either at the upper left-hand corner of the graph, or along the side, next to the vertical axis. To label along the side, right-click while the cursor is in the genome browser window. A menu will appear. Go to the Options submenu, and click the **Label vertical axis on side** option. To label at the top, go to the same menu, and choose **Label vertical axis at top**.
- **Hide Experiment Name/Show Experiment Name:** You can show or hide the experiment name (look for it in the upper right corner of the Genome browser) by right-clicking in the browser and toggling **Hide experiment name** from the Options submenu.
- **Graph raw data/Graph normalized data:** You can display raw or normalized data (as shown in the upper right corner of the Gene Inspector window) by right-clicking in the browser and toggling **Graph raw data** from the Options submenu.

Copyright 1998-2001 Silicon Genetics

The Error Bars Submenu

Before you turn the error bars on, go to **`Experiments > Change Experiment Interpretation`** and select the **Use Global Error Model** checkbox. Please refer to "Global Error Models" on page 2-26 and "Global Error Models Technical Details" on page N-1 for more details and restrictions on this topic.

- **Show Error Bars/Hide Error Bars:** You can show or hide error bars by right-clicking in the genome browser and toggling **Show error bars** from the Options submenu. Error bar will only show for averaged data, if you cannot get error bars to show, check your parameters or re-define one as a replicate.
- **Standard error bar:** This feature only works in the Graph view when the error bars are showing. You can display the Standard deviation error bars by right-clicking in the genome browser and toggling **standard deviation error bar** from the Options submenu. This feature is not enabled in the Gene Inspector window. See "Common Commands in the Experiment Specification area" on page -10 for more information.
- **Standard deviation:** This feature is only available in the Graph view when the error bars are showing. Please contact Silicon Genetics' technical service department at support@sigenetics.com or call 650-367-9600.
- **Min/Max:** This feature is only available in the Graph view when the error bars are showing. Please contact Silicon Genetics' technical service department at support@sigenetics.com or call 650-367-9600.

Common Commands in the Navigator

Right-clicking over a list or a folder will often bring up a list of commands related to that folder.

Mac Users should use Control-Click to activate pop-up menus.

- **Display:** This command will change the view to the data-object selected.
- **Inspect:** This command will bring up the Inspector window for the data-object, whether it is a list, tree or something else. Most of the fields in the History section of the Inspect window (and for some items you will have only a History section) are editable.
- **Attachments:** This command allows you to view any attachment to any data-object in the navigator. You may also add, remove or change the name of any attachment (by using the **Save As** command). Attachments can be text files, pictures, or anything you would like to have associated with a specific data-object in GeneSpring.
- **Delete:** Selecting this will result in a caution window asking you to verify the deletion of the data-object. Click **Yes**, and your data-object will be gone forever. Some data-objects cannot be deleted, you should see a pop-up window with a message to that effect.
- **Rename:** Selecting this will result in a new window asking for the new name. Type in the new name and click **`OK`**.

Copyright 1998-2001 Silicon Genetics

- **Publish to GeNet:** This will bring up the GeNet UpLoad Window. From here you can load data from this list into the GeNet database. Please see “Publish to GeNet” on page 6-6 or the *GeNet User Manual* for more details.
- **Save to disk:** This feature will save any data-object to your local drive if it is not already there. Typically, only if you are working from a server or from GeNet will this be a useful option.

The Main Folder Pop-up Menus

A right-click over a main folder (such as Gene Lists or Classifications) will produce a small menu possibly including some or all of the following:

Mac Users should use Control-Click to activate pop-up menus.

- **Use As Classification:** This command will shift your current view into classification (if you are not there already) and list the genes under each classification heading. The coloration will not change. See “Classifications View” on page 3-9 for more information.
- **Use As Coloring:** This command will change the current coloring of your view to a coloration scheme reflecting the folder chosen. The colorbar will change to a list of blocks with captions telling you which list is which. See “Color by Classification” on page 3-34 for more information.
- **Split/Unsplit Window:** This feature allows you to view multiple graphs simultaneously in the genome browser. You can also unsplit the window by selecting **`View > Unsplit window`**.
- **Publish to GeNet:** This will bring up the GeNet UpLoad Window. From here you can load data from this list into the GeNet database. Please see “Publish to GeNet” on page 6-6 or the *GeNet User Manual* for more details.
- **Clear:** The command will clear the current display.
- **Delete:** This command will delete the data-object. There will be a confirmation box.

The Gene Lists Folders Pop-up Menus

A right-click over a subfolder in the main Gene Lists folder will bring up the following commands:

- **Use As Classification:** This command will shift your current view into classification (if you are not there already) and list the genes under each classification heading. The coloration will not change. See “Classifications View” on page 3-9 for more information.
- **Use As Coloring:** This command will change the current coloring of your view to a coloration scheme reflecting the folder chosen. The colorbar will change to a list of blocks with captions telling you which list is which. See “Color by Classification” on page 3-34 for more information.

 Copyright 1998-2001 Silicon Genetics

- **Split/Unsplit Window:** This feature allows you to view multiple graphs simultaneously in the genome browser. You can also unsplit the window by selecting **`View > Unsplit window`**.

The Gene List Subfolder or Gene List Pop-up Menus

A right-click over a gene list will bring up the following commands:

- **Display List:** The number of genes displayed in the genome browser can be limited by choosing a gene list. Creating gene lists can be done in a number of different ways. For detailed descriptions of how to do this see "Filter Genes Analysis Tools" on page 4-1. The Gene Lists folder in the navigator lists all of the gene lists GeneSpring currently knows about. This includes lists you have made, and the list currently displayed in the genome browser. There are some subfolders, such as the "PIR keywords". The subfolders are marked with a plus sign next to their icons. Clicking one of the proffered gene lists (those with a DNA-on-a-page icon) selects that list to be displayed in the genome browser.
- **Translate:** The options, new in GeneSpring version 4.0 allows you to find genes in one genome that are also present in other genomes. Please refer to "Making Lists of Homologs and Orthologs" on page 4-31 for more details on this feature.
- **Display As Second List:** Depending on the view you are currently looking at this command may bring in a second list, all colored in green.
- **Venn Diagram:** This command allows you to assign various lists colors within a Venn Diagram. The submenu contains three options: left, right and bottom. See "Color by Venn Diagram" on page 3-33 for more details.
- **Use on Scatter Plot:** This option will give you two selectable items, **Vertical Axis** and **Horizontal Axis**. You can assign data from this list as one or the other.
- **Delete List:** Selecting this will result in a caution window asking you to verify the deletion of the list. Click **Yes** to delete.
- **Inspect:** This command brings up the Inspect Gene List window where you can view many details about the history and contents of your list. Please refer to "List Inspector" on page 3-44 for more details.

The Experiment Subfolder Pop-up Menus

A right-click over an experiment will bring up the following commands:

- **Display Primary Experiment:** Selecting this option will reset the genome browser to show that experiment. It is quicker to just select the experiment through the navigator with a left-click.
- **Set Secondary Experiment:** This will add the secondary experiment to the genome browser.
- **Inspect:** This will bring up a window with the administrative information associated with this experiment. You can click the **Edit** button to change most of the information presented in the Inspect window.

 Copyright 1998-2001 Silicon Genetics

- **Delete Experiment:** Selecting this will result in a caution window asking you to verify the deletion of the experiment. Click **Yes** to delete.
- **Delete Experiment Interpretation:** Selecting this will result in a caution window asking you to verify the deletion of the interpretation. Click **Yes** to delete.

The Classifications Subfolders Pop-up Menus

A right-click over a classification will bring up the following commands:

- **Set As Classification:** This command allows you to apply the classification system of that folder to whatever list your are currently viewing. Please see "Classifications View" on page 3-9 for more details.
- **Set As Coloring Scheme:** This command allows you to use a set of classifications as a coloring scheme. Each set will be assigned a color and will display in that color by GeneSpring. Please see "Color by Classification" on page 3-34 for more details.
- **Split/Unsplit Window:** This feature allows you to view multiple graphs (or any other display type) simultaneously in the genome browser. You can also unsplit the window by selecting **`View > Unsplit window`**.
- **Make Gene Lists:** With this command you can make a list of a classification. The New Gene List window will appear asking you to choose/create a folder and name your new list.
- **Inspect:** This will bring up a window with the administrative information associated with this experiment. You can click the **Edit** button to change most of the information presented in the Inspect window.

Common Commands in the Experiment Specification area

While there are no new commands available by right-clicking in the experiment specification area, there are several items you can show or hide.

- **The Series Variable:** You can change the series variable (parameters such as time or drug concentration) by moving the slider in the scroll bar at the bottom of the window. The series variable is represented by the green ConditionLine in the genome browser.
- **Animate:** This command moves the series variable forward automatically. To turn this feature on, simply click in the **Animate** checkbox in the gray box at the bottom of the browser display, or select the **`View > Animate`** checkbox menu item. If you are viewing Color By Expression, the colors will change according to the expression and trust of each data point.
- **Zoom Out Button:** This command reverses zoom-in by a factor of two in each direction. There are four ways to decrease magnification. One method is to click the **Zoom Out** button in the experiment specification area until the desired magnification is reached. Another method is to use **`View > Zoom Out`**. A third method is to right-click while the cursor is in the genome browser. Select the **`Zoom Out`** option of the resultant pop-up menu.

 Copyright 1998-2001 Silicon Genetics

- **Picture:** To remove the picture at the bottom right of the main GeneSpring window select **`View > Visible > Picture`**. The picture checkbox menu item should not have a checkmark after this operation is performed. To display the picture, go to the same menu and click in the **Picture** checkbox menu item, leaving a check in the checkbox menu item. Secondary Picture:

The secondary picture will be shown in the very bottom right corner of the GeneSpring Window.

- **Secondary Animation Controls:** The secondary animation controls are underneath the primary and behave in the same manner.
- **Magnification:** To hide the numerical magnification value and the **Zoom Out** button which appears in the bottom gray box of the browser display, select the **`View > Visible > Magnification`** checkbox menu item to deselect. The magnification checkbox menu item should not have a checkmark after this operation is performed. To display the numerical magnification value and the **Zoom Out** button at the bottom of the browser display, go to the same menu and select the Magnification checkbox, leaving a check in the checkbox menu item. This does not disable the zoom functions, which can still be done through other menus. See the Zoom In, Zoom Out, and Zoom Fully Out commands above, for a description of these functions and directions for how to employ them.

Copyright 1998-2001 Silicon Genetics

Copyright 1998-2001 Silicon Genetics

Appendix Q Glossary

A

Array. a set of spots on a chip, typically expressed as a set of intensity measurements. An array generally has one sample. If all of the interesting genes fit onto one array, the terms array, chip and sample can be considered synonymous.

Array Layout. synthetic picture of genes on arrays. The Array Layout view can be used to check for gross slide related problems

C

Chip. the measurements from a glass slide containing DNA samples for microarray analysis.

Classification. a grouping of genes by k-means or SOM clustering that is stored in the Classifications folder.

Classification View. allows you to visualize one condition or experiment by organizing the genes according to previously defined functional categories, or by some other previous knowledge of the genes. For example, of you have genes arranged into many lists in the same folder, you can use that folder to categorize the genes on screen.

Colorbar. the rectangle on the far right of the main GeneSpring screen. The intensity of the colorbar in GeneSpring indicates how reliable the data for each gene is. Indicate a raw signal strength value to be considered very reliable (a high signal strength) value, an average (a medium signal strength) value, and an unreliable (a low signal strength) value. Any gene with a signal strength (control) above the value indicated as a high signal strength will be colored using the brightest color appropriate, any gene with a signal strength below the value given for unreliable data will be almost black in color. The medium signal value gives the value for the mid-point of the color bar, and genes with a medium signal strength are colored halfway between the two color extremes.

Condition. a grouping of one or more samples.

Control. an experiment data set that provides a comparison or contrast to experimental results.

Control Strength. (see also expression strength) the quantity divided by the raw value to get the normalized value.

Cluster. a collection of genes that have been grouped according to a certain criteria, such as similar mean expression values.

D

Data Objects. any downloadable or uploadable items in GeneSpring, such as genomes, gene lists, classifications, etc.

Dendrogram. a diagram showing hierarchical relationships, based on similarity between elements, for example, similarity of gene expression levels.

Copyright 1998-2001 Silicon Genetics

Drawn Gene. lines representing gene profiles that you draw in the genome browser. You can then search for genes matching that profile.

E

Experiment. a group of conditions associated together under one name. This generally means they were all performed using a particular set of parameters.

Experimental Parameter. a variable used to describe the condition or conditions during an experiment. A set of parameter values defines a single experimental parameter. When the word "parameter" is used alone, it usually refers to an experimental parameter.

Experiment Tree. a dendrogram used to show the relationships between the expression levels of conditions.

Experiment Specification Area. the area under the genome browser that indicates which, if any, sub-experiments, is being displayed, e.g. a particular time point in a time series experiment.

Expression. production of mRNA through transcription of a DNA gene sequence.

Expression level. the amount of mRNA produced by a given gene under specific conditions.

External Program. analysis programs outside GeneSpring which can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

F

Folders. the yellow icons denoting the various directories where data is stored, e.g., Gene Lists folder, Experiments folder, etc.

G

Gene List. a list of genes based on some criteria.

Gene Tree. dendrograms used as a method of showing relationships between the expression levels of genes over a series of conditions.

Genome. the set of all genes on a chip or array.

Genome Browser. the area of a GeneSpring window containing a visual representation of genes.

I

Interpretation. Experiment Interpretations tell GeneSpring how to treat and display your experiment parameters and how normalized values should be treated.

M

Main Screen. the first GeneSpring window that appears after you open a genome, such as the default yeast genome window that appears after initially starting the program.

Measurement. the smallest "unit" of data recognized by GeneSpring. These raw values can be seen in the upper right table in the Gene Inspector.

Copyright 1998-2001 Silicon Genetics

Menu. pull-down options that allow you to perform tasks in GeneSpring. The main menu can be found at the top the main GeneSpring window (PC) or at the top of your screen (Mac).

N

Navigator. the left panel of GeneSpring windows containing data organized into folders.

Normalize. the use of statistical methods to eliminate systematic variation in microarray experiments that can influence measured gene expression levels.

P

Panel. section of a window or screen.

Pathways. A pathway is a graphical representation of the interaction between gene products in a biological system. Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context.

Parameter-Value. one of the possible values assigned to a variable. For example, in the equation:

X ={1, 2, 3 or 4}

"X" is the experimental parameter and the numbers 1, 2, 3 or 4 are each a different parameter-value of "X". A more pertinent example is the parameter values breast cancer, kidney cancer, liver cancer, brain cancer, and no cancer could all be different parameter values for the experimental parameter "cancer".

Parameters.

Color Code is similar to a discrete parameter, except you would expect points on a graph with the same parameters other than this one to be at the same horizontal position. Colors would then be typically used to distinguish these points. Typical examples are the same as for non-continuous parameters. This may be referred to as category.

Continuous Parameter is a numerical parameter for which interpolation makes sense. Graphs using this parameter are line graphs. If there are no continuous parameters in an experiment, then histograms will be shown instead of line graphs. A typical example of a continuous parameter is time, or drug concentration. Continuous parameters can optionally be made logarithmic for display purposes.

Non-continuous Parameter is a (possibly numerical) parameter for which drawing lines between points does not make sense, but you still wish to graph it along the horizontal axis. Typical examples of such parameters are drug type, strain of the organism under study, or tissue type. GeneSpring will typically display smaller graphs side by side in the genome browser. This may also be referred to as discrete.

Replicate is not interpreted by GeneSpring. Instead, it is considered a tracking identifier. Sub-experiments that have all parameters (other than the "Replicate" parameter) the same are considered repeats. These are visually represented on graphs by taking the median of the data values and plotting error bars. Typical examples of such parameters are database identifiers, and individual organism names.

Picture.

Copyright 1998-2001 Silicon Genetics

Pop-up Menu. A list of options that appears from a sub-menu or by right-clicking (Option-click for Mac).

R

Replicate. Replicates can be multiple spots on the same array representing the same gene (also referred to as a copy), the same sample in more than one array or a biological replicate - that is equivalent samples taken from more than one organism. A parameter defined as a replicate is graphically a hidden variable; no visual distinction is made based upon this parameter or its parameter values.

Regulatory Sequence. the sequence upstream of a given gene to which regulatory enzymes bind, determining the amount of expression of a particular gene.

S

Sample. the measurements taken from one or more chips containing a single liquid sample. OR the data generated from a biological object placed onto an array or set of arrays.

Slider. a horizontal scrollbar at the bottom of the GeneSpring window that changes the display of genes from one sub-experiment to another, e.g., in a time series experiment, the slider moves the displayed genes across the different time periods.

T

t-test. T-tests calculate p-values which measure the significance of differential gene expression in each condition.

Trust. a measure of reliability of the data.

Two-color experiment. an experiment where a control is used.

V

Variable. a factor such as a disease, drug concentration, patient name, pipette number, time, the strain of organism tested, or who performed the experiment, etc. These variables allow you to look for meaningful patterns in you data and deal sensibly with replicate experiments.

Copyright 1998-2001 Silicon Genetics

Index

A
adding extra genes H-4
affine background correction 2-23, G-18
All Samples to Specific Samples J-18
Animation Controls 3-6
API E-1
Array Element List. *see* Master Gene Table
Array Layout view 3-22
Array Photos D-12
Attachments P-7
B
background signal J-10
Bar Graph view 3-8
browser display
  Picture 3-7
Build Simplified Ontology 2-16
C
Calinski and Harabasz index 3-47
Change Coloration 3-31
Change correlation 4-16, L-5
Change Experiment Interpretation 2-17
change experiment name 3-42
Change Vertical Axis Range P-5
changing restrictions 4-9
Class Predictor 5-15
Classification Inspector 3-46
  class 3-47
Classification view 3-9, 3-27
CLI E-2
Cluster P-4
  results 5-11
Cluster Menu. *see* Tools Menu
Clustering window
  similarity definitions L-1
Color
  by Classification 3-34
  by Parameter 2-14, 3-33
  by Secondary Experiment 3-35
  by Significance 3-33
  by Venn Diagram 3-33
  changing the defaults B-2
  No Color 3-34
  Trust 3-32
Color by Primary Experiment. *see* Color by Expression
color code parameter J-3
Colorbar J-19
Common Name H-2
Compare Genes to Genes view 3-24
  Interesting Genes 4-21
complementary bases
  show/hide P-6
Complex Correlations 4-18
Condition Inspector 3-43
Conjectured Regulatory Sequence 4-29
constant value. *see* hard number
continuous parameter J-3
Control Channel Background Column D-11, J-11
Control Channel Values D-11, J-11, J-15
  minimum value J-15
  pre-normalized data J-16
Copy lists to clipboard 3-46
Copying and Pasting data F-1
correlation
  weighted 5-2, 5-11
Correlation commands 4-14, L-2
Correlation Equations
  Change correlation 4-16, L-5
  Distance 4-17, L-4
  Pearson correlation 4-17, L-2
  Smooth correlation 4-16, L-4
  Spearman Confidence 4-17, L-3
  Spearman correlation 4-17, L-3
  Standard correlation 4-16, L-2
  Two-sided Spearman Confidence 4-17, L-3
  Upregulated correlation 4-16, L-5
D
Data Column Location D-10, J-9
data directory H-6, K-8
Data File Format D-4
Data File Header Lines D-8, J-7
Data Import Wizard
  Experiment D-3
  Genome C-1
data location K-8
data objects 6-6
Database E-1
  JDBC driver B-1

Copyright 1998-2001 Silicon Genetics

DBMS E-1
dendrogram. *see* Tree View
Describe your Data Files D-6, J-6
Display Parameters J-2
Distance 4-17, L-4
Downregulated Color B-2
E
Each Gene to Itself J-18
  minimum average J-18
Each Sample to Itself J-17
  minimum average J-17
EC Number H-2
Edit Menu P-2
equations
  overall correlation 5-3
Error bars P-7
Euclidian metric L-4
Experiment Inspector 3-41
  buttons 3-43
  interpretations 3-42
  normalizations 3-42
  notes 3-42
  parameters 3-42
experiment installation files K-1
experiment interpretation
  changing 2-17
  Fold change 2-19
  log ratio 2-18
  vertical axis 2-18
Experiment Name J-1, P-6
experiment parameter 2-11
  condition 2-13
  multiple 2-12
  parameter-value 2-11
Experiment Wizard D-3
experimental data file K-1
explained variability 3-47
Export data
  by copying F-4
  to External Program interface 4-40
  to GeNet 6-6
expression values
  determining G-1
External Program interface 4-40
F
FAQ A-1
File Menu P-2
files
  .database E-4
  .experiment J-1
  .gbk C-2
  .homology 4-31
  .layout M-2
  .seq C-3
  FileAccess.jar 4-44
Filter Genes
  Condition to Condition Comparison Restriction 4-7
  Data File Restriction 4-7, 4-8
  Expression Percentage Restriction 4-3
  Expression Restriction 4-7
  removing restrictions 4-9
  restricting data types 4-8
Find Gene 3-4, P-2
Find Potential Regulatory Sequence 4-26
Find Similar Genes 3-40
Finish D-16
Flags D-11, G-17, J-12
formula notation L-1
Functional Classification 3-27
  clear or remove 3-28
G
GATC E-2
GenBank Accession Number H-3
Gene Inspector 3-37
  Control 3-39
  Correlation Commands 4-14
  Description 3-39
  Normalized 3-39
  notes 3-40
  Raw 3-39
  Save Profile 3-40
  Student's t-test 3-39
  t-test p-value 3-39
  Web Connections 3-40
Gene Name D-9, J-8
Gene Name Prefix Removal D-9, J-8
Gene Name Suffix Removal D-10, J-8
gene similarity L-1

Copyright 1998-2001 Silicon Genetics

GeneSpider 2-15, P-4
lists from annotations 4-19
GeneSpring Basics Instructional Manual A-1
GeneSpring User Manual A-1
GeNet 6-6
GeNet Database A-2
Genome Browser
printing 6-2
Genome Browser. *see also* Browser display
Graph by Genes view 3-26
commands 3-26
Graph raw data P-6
Graph view 3-7
color by secondary experiment 3-35
Graphics Specifications D-15
Guess the rest D-11
H
hard number G-7
headlines J-7
Help Menu
About A-2
FAQ A-1
Manual A-1
SiG on the Web A-2
System Monitor A-2
Version Notes A-1
Hide All 3-6
Hierarchical Clustering View. *see* Tree View
homologous genes 4-31
Horizontal Label P-6
housekeeping genes 2-22
How to Display the Parameters D-5
I
Import data
by pasting F-1
from GeNet 6-8
Inspectors
Condition 3-43
Experiment 3-41
Gene 3-37
Interpretation 3-41
installation files K-1
installing GeneSpring 1-1
Interpretation Inspector 3-41
interpretations 2-17
J
JDBC driver B-1
K
KEGG 4-25
Keywords H-3
K-means clustering 5-9
Maximum Iterations 5-11
Number of Clusters 5-11
Kyoto Encyclopedia of Genes and Genomes 4-25
L
layout file K-2
negative controls J-15
positive controls J-16
region specifications J-9
List Inspector 3-44
Lists
Find Interesting Genes 4-21
Find Similar 4-13
from annotations 4-19
p-value 4-11
Regulatory Sequences 4-29
Venn Diagram 4-19
Load Sequence P-5
command 3-13
M
Magnification 3-6
Main GeneSpring Screen. *see* Browser display
Make New Tree 5-1
Mapped format K-7
Common Name H-2
custom H-3
EC Number H-2
function H-3
GenBank Accession Number H-3
gene list formats H-2
Keywords H-3
Map H-2
phenotype H-3
Protein Product H-3
Public Medline accession number H-3
sequence H-3
Systematic Name H-2
Mapping information H-2
Master Gene Table 2-15, C-3, H-1
gene list formats H-1
mathematical notation L-1
measurement flags D-11, G-17, J-12
Abs/Call 2-17

memory 1-2
Minimum Distance 5-3, 5-4
missing expression values L-1
mock phylogenetic 5-2
Multi-Experiment Correlation 4-14
N
name function H-2
    gene list formats H-2
name list H-1
    gene list formats H-1
Navigator 3-6
negative control strengths G-18
Negative Controls J-14
new Pathway 4-24
nodes 5-12
non-continuous parameter J-4
normalization options 2-21
    All Samples to a Specific Sample D-15
    All Samples to Specific Samples G-10
    all samples to specific samples 2-25
    background subtraction 2-21
    constant value 2-24
    Control Channel Values D-13
    Control Channel Values for Each Gene G-3
    Distribution of All Genes G-6
    distribution of all genes 2-23
    Each Gene to Itself D-15, G-8
    Each Sample to a Hard Number D-14, G-7
    Each Sample to Itself D-14, G-6
    gene to itself 2-25
    Global Scaling G-6
    hard number 2-24
    Negative Controls 2-21, D-13, G-2
    order 2-21
    per chip 2-22
    per spot 2-22
    positive control 2-22
    Positive Controls D-13, G-5
    pre-normalized data 2-24
    Region Normalization G-15
normalization techniques G-1
Normalization to Specific Samples G-10
Number of Arrays D-4, J-1
Number of Parameters D-5, J-2
O
ODBC E-1
one-color experiments 3-32
opening new genomes 1-17
Options
    Change Vertical Axis Range P-5
Ordered List view
    Interesting Genes 4-21
ORF direction
    Ignore P-6
    Show P-6
orthologous genes 4-31
over-expressed color
    changing B-2
P
Panning 3-1
parameter
    numeric F-2
Parameter Characteristics D-5, J-2
Parameter Interpretations
    fold change (+100% is 1,-50% is -1) 2-19
    log ratio 2-18
    ratio 2-18
    ratio of signal/control 2-18
Parameter names J-2
Parameter Values D-5, J-4
Parameters
    category J-3
    color code J-3
    continuous J-3
    discrete J-4
    display D-5
    display instructions J-2
    non-continuous J-4
    non-numeric 2-10, 2-13, F-2
    numbers J-2
    numeric 2-10, 2-13
    order 2-10
    replicate J-4
    set J-4
    units J-2
Pass Fail column. *see* Flags
pasting data D-3
Pathway view 3-23, 4-23
    adding new elements 4-24
    multiple genes 4-24
PCA. *see* Principal Components Analysis
Pearson correlation 4-17, L-2
Percent Explained variability 3-47
phase offset 4-18
Phenotype H-3
phylogenetic tree. *see* Tree View

Copyright 1998-2001 Silicon Genetics

Physical Position view 3-10
commands 3-13
Picture 3-6
Pictures J-13
Positive Controls J-16
minimum average J-17
Predictor 5-15
Preferences window B-1
background color B-3
color B-2
data directory B-1
Database B-1
Default Correlation B-5
Default Font B-5
default genome B-1
Desired Memory B-5
Disk Cache Size B-5
firewall B-4
GeNet Address B-5
License Manager B-5
Restrict Gene List Searches B-5
selected color B-3
structure color B-3
Unique ID prefix B-5
web browser defaults B-4
Principal Components Analysis 5-5, P-4
Print
List 3-46
Printing Pictures 6-2
Trees with labels 3-18
Properties of Experiment D-4, J-1
Protein Product H-3
Publish to GeNet 6-7
P-value 4-11
R
raw data K-1
References Values. *see* Control Channel Values
region designation file K-6
Region Normalization D-8, G-15
multiple arrays J-9
Regulatory Sequence 4-26
Expected 4-28
Observed 4-28
P-value 4-28
Random Rate 4-28
Sequence 4-28
Single P 4-28
Tests 4-28
rename gene list 3-46
replicate parameter J-4
restrict data types
Control Signal 4-7
Normalized Data 4-7
Number of Replicates 4-7
Range of Normalized Data 4-8
Raw Data 4-7
Standard Deviation 4-8
Standard Error 4-8
T-test probability 4-8
restricting data types 4-8
RT- PCR Experiments D-12
S
Sample Photos D-11, J-13
Save List 3-46
Scatter Plot view 3-15
color by secondary experiment 3-35
Scripts 4-32
Secondary Animation Controls 3-6
Secondary Picture 3-6
select a gene(s) 3-4, P-1
deselect a gene 4-22
Self-Organizing Maps P-4
Separation ratio 5-3, 5-4
SGD H-2
gene list formats H-2
Show All 3-6
Show complementary bases P-6
similarity definitions. *see also* correlations
Smooth correlation 4-16, L-4
SOM
Euclidean distance 5-13
Spearman Confidence 4-17, L-3
Spearman correlation 4-17, L-3
Split Window 3-30
classification 3-35
SQL E-2
Standard correlation 4-16, L-2
Standard deviation error bar P-7
Syntax G-10
Systematic Name H-2

T
Table of Genes *see* Master Gene Table
Tools Menu P-4
Translate 4-31
translation table 4-31
Tree View 3-17
Trees
comparing genes in nodes 3-18
labels 3-18
Minimum Distance 5-3
Separation ratio 5-3
viewing 3-17
troubleshooting
Java Virtual Memory 1-2
Trust 3-32
t-test 3-39
Tutorial A-1
two-color experiments 3-32
Two-sided Spearman Confidence 4-17, L-3
U
under-expressed color
changing B-2
Update annotations 2-15
Update genes. *see* GeneSpider
Update GeneSpring A-2
upload to GeNet 6-7
Upregulated Color B-2
Upregulated correlation 4-16, L-5
Use list as Classification 3-27
V
Venn Diagram 3-33
Version Notes A-1
vertical axis P-6
Vertical Label P-6
view gene details 3-37
View Menu P-3
Array Layout 3-22
Bar Graph 3-8
Classification 3-9
Compare Genes to Genes 3-24
Graph 3-7
Graph by Genes 3-26
Pathway 3-23
Physical Position 3-10
Scatter Plot 3-15
W
Web Connections 3-40
web databases C-4
special character C-4
Welcome panel D-3
Wizard Panels
Array Photos D-12
changing panels manually D-3
Control Channel Values D-11, D-13
Data Column Location D-10
Data File Format D-4
Data File Header Lines D-8
Describe your Data Files D-6
Finish D-16
Flags D-11
Gene Name D-9
Gene Name Prefix Removal D-9
Gene Name Suffix Removal D-10
Graphics Specifications D-15
How to Display the Parameters D-5
Normalizations by All Samples to a Specific Sample D-15
Normalizations by Each Gene to Itself D-15
Normalizations by Each Sample to Itself D-14
Normalizations by Negative Controls D-13
Normalizations by Positive Controls D-13
Normalizations Each Sample to a Hard Number D-14
Number of Arrays D-4
Number of Parameters D-5
Parameter Characteristics D-5
Parameter Values D-5
Properties of Experiment D-4
Region Normalization D-8
RT- PCR Experiments D-12
Sample Photos D-11
Welcome D-3
Y
y-axis J-19
Z
zoom out P-5

Copyright 1998-2001 Silicon Genetics